European Journal for Philosophy of Science (2023) 13:18
https://doi.org/10.1007/s13194-023-00520-1

PAPER IN THE PHILOSOPHY OF THE SOCIAL SCIENCES AND HUMANITIES

Extrapolating from experiments, confidently

Donal Khosrowi[1]

Received: 4 March 2022 / Accepted: 21 February 2023 / Published online: 22 March 2023

Abstract
Extrapolating causal effects from experiments to novel populations is a common practice in evidence-based-policy, development economics and other social science areas. Drawing on experimental evidence of policy effectiveness, analysts aim to predict the effects of policies in new populations, which might differ importantly from experimental populations. Existing approaches made progress in articulating the sorts of similarities one needs to assume to enable such inferences. It is also recognized, however, that many of these assumptions will remain surrounded by significant uncertainty in practice. Unfortunately, the existing literature says little on how analysts may articulate and manage these uncertainties. This paper aims to make progress on these issues. First, it considers several existing ideas that bear on issues of uncertainty, elaborates the challenges they face, and extracts some useful rationales. Second, it outlines a novel approach, called the *support graph approach*, that builds on these rationales and allows analysts to articulate and manage uncertainty in extrapolation in a systematic and unified way.

Keywords Extrapolation · External validity · Causal inference · Uncertainty · Confidence · Bayesian evidence amalgamation · Evidence-based policy · Economics

1 Introduction

In evidence-based policy, development economics and other social science fields, researchers often conduct randomized controlled trials (RCTs) to measure the effects of a policy or other intervention. With the results established, analysts often

This article belongs to the Topical Collection: EPSA21: Selected papers from the biennial conference in Turin
Guest Editors: A.C. Love, C. Marchionni, M. Redei, J. Williamson

✉ Donal Khosrowi
donal.khosrowi@philos.uni-hannover.de

[1] Leibniz Universität Hannover Institute of Philosophy, Hannover, Germany

endeavour to *extrapolate* these effects to new target populations. It is widely recognized that extrapolation involves significant epistemic challenges because the populations where interventions are studied and target populations of interest may differ significantly (Vivalt, 2020; Cartwright, 2013a; Reiss, 2019; Steel, 2009). An important aim in making an inference to a new environment is hence to clarify whether populations are sufficiently similar, and to account for differences between them (Cartwright, 2013b; Khosrowi, 2019). A growing literature articulates what assumptions, exactly, one must entertain to enable valid extrapolative inferences (e.g. in what respects two populations must be similar), and what challenges are involved in supporting these assumptions (Hotz et al., 2005; Bareinboim & Pearl, 2012, 2016; Cartwright, 2013b; Muller, 2014, 2015, External validity, causal interaction and randomised trials: The case of economics, unpublished manuscript; Athey & Imbens, 2017; Duflo, 2018). Yet, while it is also widely recognized that real-world cases invariably involve substantial uncertainties regarding such assumptions, there is a lack of concrete proposals for how to articulate and manage such uncertainties. This is surprising since high-stakes decision-making that relies on extrapolation, e.g. implementing large-scale policies in new environments, involves a pressing need to understand and mitigate these uncertainties.

This paper pursues two aims relevant to addressing this need. The first is to consider how existing approaches, including Bayesian approaches for evidence amalgamation (Landes et al., 2018), may help with articulating uncertainty in extrapolation, and to explore what challenges they face. I argue that while existing approaches help us understand how to bring varied evidence to bear on *specific* assumptions, they struggle with telling us how the support for these assumptions compounds and propagates onto an overall conclusion[1]. My second aim is to sketch a novel approach, called the *support-graph approach* (SGA), that builds on existing resources to facilitate more comprehensive assessments of uncertainty. At its center, SGA draws on a prominent methodological rationale found in sensitivity/robustness analysis, error statistics, risk analysis, and other places (Rosenbaum, 2002; Mayo & Spanos, 2004; Roy & Oberkampf, 2011): to understand a state of uncertainty surrounding a prediction, one should consider what would happen if one's assumptions turned out false. SGA systematizes this rationale in the context of extrapolation and thus helps analysts and decision-makers better articulate, understand, and ameliorate the uncertainties they face.

Section 2 outlines problems of extrapolation in social science contexts, the uncertainties arising there, and how they matter to decision-makers. Section 3 considers existing approaches and discusses their limitations. Section 4 sketches the support-graph approach for articulating and managing uncertainty. Section 5 considers and responds to some concerns about SGA. Section 6 concludes.

[1] I use "conclusion" and "prediction" interchangeably to refer to what an analyst concludes/predicts about a target.

2 Extrapolation and uncertainty

The last decade has seen significant progress in developing strategies for extrapolating causal effects from social science experiments and quasi-experimental studies (Cartwright, 2013b; Cartwright & Hardie, 2012; Athey & Imbens, 2017; van Eersel et al., 2019; Duflo, 2018). Amongst the most flexible are approaches that allow analysts to adjust for *quantitative* differences between populations. For instance, if the strength of an effect depends on a variable Z, these methods help us estimate the effect conditionally on Z, and form a prediction regarding a target population B that adjusts for differences in Z between populations (Hotz et al., 2005; Crump et al., 2008; Bareinboim & Pearl, 2012, 2016). Things are more difficult when populations exhibit deeper, *structural* differences, i.e. they differ not only in the values or distributions of variables, but also regarding the structure of the causal mechanisms that underlie the effects of interest. Here, even sophisticated approaches can only licence inferences to the extent that a wide range of structural similarities can be assumed (Muller, 2014, 2015, External validity, causal interaction and randomised trials: The case of economics, unpublished manuscript; Hyttinen et al., 2015; Khosrowi, 2019).

To illustrate, suppose we have estimated the effects of providing poor households with access to microcredit in a large-scale RCT in population A. Suppose the intervention has shown significant positive effects by helping individuals make investments to start small businesses, which subsequently increases household income and welfare. Consider now an analyst who is tasked with predicting whether microfinance will be effective in a novel population B. In making an inference about B, she has to consider whether A and B are sufficiently similar, identify where they (likely) differ, and accommodate relevant differences in her inference. For instance, individuals in B might have less entrepreneurial ability than in A (quantitative difference). Or there might be more significant practical obstacles to starting a small business in B, such as uncooperative bureaucrats (structural difference). Both kinds of differences could hamper the effects of microfinance.[2] Let us assume that our analyst endeavours to make a prediction about B, making various assumptions about their causal makeup and similarities between them (e.g. that there are no significant bureaucratic obstacles to starting a small business in B), and accommodating information about how A and B differ (e.g. concerning entrepreneurial ability). A key question she faces is: how confident can she be in her prediction, given that there might be substantial uncertainties regarding some of the similarities she has to assume?

In many evidence-based policy scenarios, it is common to make decisions based on *systematic reviews* and *meta-analyses* that provide pooled and weighted estimates from several studies with an associated uncertainty (Deeks et al., 2022). But this is not the information that our analyst needs as it only considers uncertainties intrinsic

[2] My distinction between quantitative and structural differences is not exhaustive. For instance, one might additionally consider structural differences occurring with respect to the larger, exogenous causal environment in which a given phenomenon is situated, e.g. whether a phenomenon occurs in a free market economy or not. I am open to further or differently-grained distinctions, but concentrate on the one provided here for reasons of simplicity. I thank an anonymous reviewer for highlighting this concern.

to the amalgamated studies themselves, but not those involved in reasoning to a novel population. Instead, our analyst's overall confidence in her prediction should be a function of the support that the assumptions involved in her inference enjoy. These assumptions can be articulated at different levels of detail. For the purposes of this paper, I assume that it is typically useful to fine-grain one's assumptions. So rather than supporting a blanket assumption such as "the causal mechanisms in A and B are similar", our analyst should break this assumption down into smaller *component assumptions* that each pertain to specific causal relationships or features, such as that it is similarly easy to register a small business in both populations, or that the distribution of entrepreneurial ability is similar.

My focus here is on the uncertainty that will surround these assumptions. While quantitative similarities are usually easier to support, e.g. that the distribution of a variable is similar in A and B, structural similarities, e.g. that formal and informal institutions that govern or constrain individuals' behaviors are similar, are often significantly more difficult to support.[3] So how can our analyst tell how these uncertainties bear on the confidence she may have in a conclusion? Let me consider some existing ideas and approaches that help address this question.

3 A toolbox half full

In handling uncertainty in extrapolation, we need to distinguish two questions. The question of *assumption uncertainty* asks: how confident can we be in specific assumptions needed for an inference? The question of *conclusion uncertainty* asks: how does the support for these assumptions compound and propagate onto the confidence we are entitled to have in a conclusion? I argue that, disappointingly, existing approaches do not tell us quite enough to answer both questions. But, more optimistically, I argue that we are lucky to have a toolbox half full, which contains several useful ingredients. Taking stock of what we have, I first consider a useful principle suggested by Cartwright and Stegenga (2011), and propose some refinements to it. I then consider two additional resources for making further progress. First, a Bayesian Networks approach (Landes et al., 2018) for amalgamating evidence, which can help us address the question of assumption uncertainty but not conclusion uncertainty. Second, the Confidence Approach employed by Roussos et al. (2021) in the context of uncertainty management in climate modelling, which provides a useful rationale that gets us closer to what we need.

[3] The distinction between quantitative and structural features/similarities is neither definitive nor fully sharp, e.g. a difference in the value of a variable distribution or parameter can often be recast as, or explained as the result of, a lower-level or extraneous structural difference. My view here prioritizes epistemic issues and insists that, given a specific level of causal abstraction and isolation determined to be appropriate for an inquiry, we can nevertheless meaningfully distinguish between quantitative differences (e.g. values/distributions of variables and parameters) and structural differences (e.g. whether or not relationships exist between pairs of variables), and treat those latter ones as primitive without committing to the idea that there aren't richer causal stories to be told.

3.1 The weakest link

As Cartwright and others emphasise, even high-quality evidence indicating a causal effect somewhere is often a poor guide, all by itself, to what will happen elsewhere (Cartwright, 2011, 2013a; Cartwright & Stegenga, 2011; Cartwright & Hardie, 2012). For that, an inference is needed, and this inference will involve substantive causal assumptions that require support. Things get complicated, however, when we assess conclusion uncertainty: how does the support for these assumptions underwrite our overall conclusion? Cartwright and Stegenga (2011) argue that:

> … [a] chain of defense for the effectiveness of a policy, like a towing chain, is only as strong as its weakest link. So the investment in rigour for one link while the others are left to chance is apt to be a waste. To build the entire chain one may have to ignore some issues or make heroic assumptions about them. But that should dramatically weaken the degree of confidence in the final assessment. Rigour isn't contagious from link to link. If you want a reasonably secure conclusion coming out, you'd better be careful that each premise is secure enough going in. (2011, 293).

The main purpose of this passage is to emphasise that the putatively high credibility of RCTs does not warrant high confidence in conclusions about novel populations, at least not unless the assumptions required for such conclusions are well-supported. I agree with this view. Here, I focus attention on how Cartwright and Stegenga suggest that the support for these assumptions bears on the overall confidence in a conclusion. Let me recast the essence of their point as the following principle, which provides a useful starting point to recognize additional complexities:

Weakest Link (WL) The confidence we are entitled to have in a conclusion may only be as high as the confidence we have in the assumption that is least well supported.

Let us consider an example to see how WL might work. Suppose we need to assume that a causal mechanism in population A, $X \rightarrow Z \rightarrow Y$, is similarly instantiated in B. We can subdivide this into two component assumptions:

$A1$: there is a causal relationship $X \longrightarrow Z$ in B
$A2$: there is a causal relationship $Z \longrightarrow Y$ in B

What does WL tell us about how the support for these assumptions bears on the confidence we have in a conclusion about B? There are at least two ways to understand WL formally. Let $Con_C \in [0,1]$ be the confidence we have in a conclusion C about population B. Let $Con_{A1}, Con_{A2} \in [0,1]$ respectively be the confidence we have in our two causal assumptions, $A1$ and $A2$. One way of rendering WL is as a min-function of the support for $A1$ and $A2$, formally, $Con_C = min[Con_{A1}, Con_{A2}]$.[4] To

[4] A more general rendition would be to say that Con_C should be less or equal (rather than equal) to the confidence we have in the least well supported assumption. I only consider the stricter rendition in terms of equality here as this makes the numerical examples easier to parse, but note that I am sympathetic to the more general rendition, too.

take an example, if $A1$ is strongly supported and $A2$ weakly, say $Con_{A1} = 0.9$ and $Con_{A2} = 0.1$, then the confidence in our conclusion may only be as high as our confidence in $A2$, i.e. $Con_C = 0.1$.

An alternative rendition of WL is to think of Con_C as a multiplicative function, following the product rule for computing joint probabilities of independent events[5]. Formally, $Con_C = Con_{A1} * Con_{A2}$. So, in our example with $Con_{A1} = 0.9$ and $Con_{A2} = 0.1$, this rendition would return a lower value for the confidence in our conclusion: $Con_C = 0.09$. In both cases, Con_C cannot be higher, and might indeed be lower, than the confidence we place in the assumption that is least well supported.

I take it that the first rendition of WL is not plausible, since Con_C would remain unresponsive to potentially large changes in the support for a body of assumptions. Take two weakly supported assumptions, $Con_{A1} = 0.1$ and $Con_{A2} = 0.11$ and now consider what happens if the support for $A2$ were to become much stronger, e.g. we learn new evidence that pushes the support to $Con_{A2} = 0.99$. The min-function rendition would return the same values for $Con_C = 0.1$ in both cases, which seems implausible as it fails to track that there is now considerably less uncertainty regarding one of two crucial assumptions. The multiplicative rendition, by contrast, accounts for this change by moving us from $Con_C = 0.011$ to 0.099.

While intuitively more compelling, the multiplicative rendition also rests on a crucial but not always plausible assumption, namely that our assumptions are logically, probabilistically, and causally independent. This assumption is plausible in the contexts where WL originated and extrapolation is considered to proceed in terms of arguments with premises (Cartwright, 2013b). Here, the premises at stake, e.g. P1: "the causal mechanisms in two populations are sufficiently similar" and P2: "important variables and parameters have similar distributions in both populations" may often plausibly be thought of as independent affairs. Why should we think, for instance, that the age distribution being similar between two populations (P2) also makes it more likely that the causal mechanisms governing a social policy's effectiveness would be similar between them (P1)?

However, once we unpack general premises like "causal mechanisms in two populations are similar", additional complexities arise. Here, analysts will often want to investigate and assert similarities regarding more specific, component features of the mechanisms at issue, such as specific causal relationships or other relevant causal features. When zooming in on such features, the relationships between our assumptions, the evidence supporting these assumptions, and our overall conclusion about a target can become more complicated, undermining the independence requirement of WL. Let me discuss some cases and explain how they can help us refine WL.

3.1.1 Case 1: Varying causal relevance

Suppose we want to predict the effects of gating alleyways on the incidence of burglaries (Sidebottom et al., 2018; discussed in Cowen & Cartwright, 2019;

[5] I thank an anonymous referee for suggesting this second reading.

Khosrowi, 2022). Evidence from A might indicate that alley gates are effective in decreasing burglaries there. But our target B might importantly differ. Suppose there are two causal features (support factors, in the language of Cartwright, 2013a) that matter for how well alley gates prevent burglaries:

$F1$: people are willing to lock the gates.
$F2$: gated areas are well-lit at night.

Both features are relevant, but to different degrees. If gates are not locked, they simply fail to establish any physical barrier in the way of burglars entering homes. If gated back alleys are not well-lit, this is mostly fine, but every now and then some burglars might still take their chances and climb the gates. How confident can we be in the conclusion that alley gates will be effective in B, given evidence speaking to $F1$ and $F2$ there? If we have little support (say, at levelα) for $F1$ while $F2$ is strongly supported (at level β), this conclusion seems shaky. Light deters, but not as much as locked gates. But if we have little support (α) for $F2$ while $F1$ is strongly supported (β), our conclusion seems much more secure. The important difference is that $F1$ is causally more relevant to our conclusion, so the same degree of support in its favor weighs more heavily on the confidence our conclusion enjoys, and WL should recognize this nuance.

3.1.2 Case 2: Varying inferential relevance

Consider a causal mechanism M taking the following form in population A:

$$X \rightarrow Z \rightarrow G \rightarrow Y$$

As before, suppose we aim to ensure that this mechanism is identical in B. To do so, we may seek to independently support three assumptions, each asserting that one of the three causal relationships comprising M is realized in B, i.e.

$A1 : X \rightarrow Z$ holds in B
$A2 : Z \rightarrow G$ holds in B
$A3 : G \rightarrow Y$ holds in B

Let us also assume, however, that the relationship $G \rightarrow Y$ is special because it only obtains as part of an M-type mechanism. If this is the case, we may have an inferential shortcut available:[6] there is no need to support $A1$, $A2$, and $A3$ to the same degree to reach a specific, high level of confidence in our conclusion. Instead, we can focus our efforts on $A3$, which then provides *indirect* support for $A1$ and $A2$

[6] This is similar to the cases that Steel (2009, 113) considers, where establishing downstream causal similarities indicates that populations are either similar upstream or, if they are not, upstream differences do not matter. Here, I envision a somewhat different case, where similarities at some stages of a mechanism are (strong) evidence for the presence of a whole mechanism-type (including at down- or up-stream stages).

as well. This does not square up well with our multiplicative rendition of WL, which fails to capture that if $A3$ is strongly supported, then regardless of the direct support for $A1$ and $A2$, the support for $A3$ sets the tone for the confidence in our conclusion because it has special inferential relevance. This is not to suggest that WL is an implausible principle, but only to highlight that WL must account for indirect support as well.[7] More generally, we can recognize two important insights: First, background (causal) knowledge is crucial for telling us how different assumptions hang together causally, and hence which of these assumptions are more important *inferentially*.[8] Second, direct evidence for one assumption can offer indirect support for another. Here, I have considered a case where, due to the nature of a mechanism, there are causal reasons for why some features (and assumptions) are more important than others. But there can also be *non-causal* reasons for why some assumptions are especially inferentially relevant. Let me expand further on this second point.

3.1.3 Case 3: Evidential interdependence

There is a broad range of cases where assumptions can be *evidentially interdependent*, i.e. where direct evidence for one assumption also confers indirect support on others. Consider an inference that rests on ten assumptions, $A1, \ldots, A10$, where the first nine are strongly supported, say 0.99, while $A10$ remains poorly supported, say 0.1. Here, WL would yield that $Con_C = 0.99^9 * 0.1 \cong 0.091$. When the causal features to which our assumptions pertain are probabilistically (and causally, and logically) independent, WL's assessment is just right: if all our assumptions need to be satisfied and A10 enjoys very little support, then A10 simply is the deal-breaker for our inference. But in other cases, independence may seem less plausible, and support might compound more strongly.

To see this, consider another case where there exists a known mechanism-type M in an experimental population. As before, assume that our assumptions A1,..., A10 encode that a range of specific features of M are instantiated in B. But suppose now that we also have reasons to believe that M's constituent causal features do not usually appear independently in the wild. In such a case, the presence of the features encoded by A1 through A9 may offer strong, indirect support for the presence of the features encoded by A10. More generally, there are often only so many ways of how the causal underpinnings of social phenomena can realistically look like; and not nearly as many as suggested by combinatorially exhausting all the ways in which a given set of causal features may be realized in a population. For instance, it may seem unlikely that there exists a concrete and independently occurring causal relationship between individuals' exposure to information about stock prices and

[7] An anonymous reviewer suggested that the multiplicative rendition could capture indirect support from the start by using conditional probabilities/confidences, e.g. $\left(Con_{A2}|Con_{A1}\right) * Con_{A1}$ to capture the indirect support for A2 afforded by A1. I agree that this is possible (and desirable, as I make clear shortly), but I do not think this is how Cartwright and Stegenga envisioned WL when emphasising that we should ensure that „[…] *each* premise is secure enough going in" (2011, 293, emphasis added) rather than taking a holistic view on what the support for a whole bundle of premises means for our confidence in a conclusion.

[8] Though there can also be non-causal reasons for why some assumptions are more important inferentially than others – I say more on this later.

their economic behaviors in any population that isn't home to a good share of people invested in stocks. Likewise, if such a relationship does exist, this can indicate that other, cognate relationships also exist (or do not exist) in that domain, such as that people's intertemporal decision-making is responsive to interest rates, or that their lifespans are not significantly affected by their hunting skills. While it is often metaphysically and methodologically sensible or convenient to consider causal mechanisms to be modular[9] (see Steel, 2009), this does not imply that any specific causal feature's probability of occurring is independent of that of others. In a causally more tightly-knit world, we can have strong enough background knowledge and theories to significantly constrain the range of possible causal arrangements in a target, indicating that some features are likely to co-occur with others or that some disparities are unlikely, given certain similarities[10]. In such a world, independence of causal features and assumptions is not generally plausible, and evidence for a set of assumptions can *mount*, i.e. compound more strongly than the product rule for independent events suggests. Importantly, in such settings we may often not have *any* direct evidence speaking to a specific assumption (e.g. A10). Nevertheless, we might still be highly confident in a conclusion, to the extent that sufficient indirect support is afforded by the compounding evidence we have brought to bear on other assumptions.

What the above considerations suggest for WL, then, is that it can remain a plausible principle, but only to the extent that it tracks an *all-things-considered* notion of support for our assumptions, direct and indirect, and that the support propagating from a set of assumptions onto a conclusion is appropriately weighted by the causal and inferential relevance of these assumptions. Yet, while WL can be refined to accommodate these insights, it still remains a general and cautionary principle. It tells us *that* the confidence we may have in a conclusion can be greatly diminished if some assumptions remain poorly supported, but it does not, by itself, offer a concrete procedure to express *how* support for specific assumptions compounds and propagates onto a conclusion. Let me turn to consider two approaches that get us closer to an account that can accomplish this.

3.2 Bayesian evidence amalgamation

Landes et al. (2018) offer a Bayesian framework for amalgamating evidence, including different kinds (e.g. statistical and mechanistic), to assess causal hypotheses about drug efficacy and harms. In doing so, they build on Bovens and Hartmann's (2003) Bayesian networks approach to modelling scientific inference. A Bayesian network is a directed graph G that encodes evidential relationships between causal hypotheses, their observable consequences, evidence regarding whether these

[9] I understand 'modularity' here as saying that, in principle, specific causal relationships could be changed without thereby affecting others.

[10] This is similar to Steel (2009), who considers the role of background knowledge in identifying stages at which mechanisms are likely to differ between populations.

consequences obtain, and information about the reliability and relevance of that evidence. Figure 1 illustrates.

Here, causal hypotheses (*C*) imply causal indications (*Ind*) that may be borne out by evidential reports (*Rep*). These reports, in turn, are modulated by nodes encoding their reliability (*Rel*) and relevance (*Rlv*). On this view, evidential support travels 'upwards' from modulated reports onto a hypothesis (Poellinger, 2020, 111, 123). An investigator interested in bringing evidence to bear on a hypothesis begins with a prior probability *P* over the variables, constrained by the conditional independencies encoded in the graph. As novel evidence becomes available, she can compute a posterior probability for the hypothesis according to the rules of Bayesian inference (Landes et al., 2018, 25) and drawing on information from a *conditional probability table* or *distribution* that expresses the conditional probabilities of nodes over the states that their parents can assume (Bovens & Hartmann 2003, 31–32).

The Bayesian approach[11] for evidence amalgamation proposed by Landes et al. offers useful resources to address the question of *assumption uncertainty*: how does evidence bear on specific assumptions involved in an extrapolation? One important virtue is that it explicitly considers issues surrounding the *weight of evidence* (see Peirce, 1878; Keynes, 1921; Good, 1985; Williamson, 2020): it is one thing to have a probabilistic belief pertaining to a hypothesis, and quite another to have an idea of how strongly supported this belief is by the evidence involved in obtaining it. The latter is often thought to be a question of the quantity and quality of evidence, as well as its diversity or consistency (Weed, 2005). Assessments of evidence weight are important in extrapolation as analysts will typically not only be interested in first-order probabilities pertaining to causal assumptions, but also in higher-order assessments of how confident they may be in these probabilities and whether this confidence is sufficiently high to licence action.[12] The Bayesian approach makes important progress towards accommodating such considerations; partly through reliability nodes that express how credible specific evidential reports are and partly through conditional probability distributions/tables, which encode ideas about how different lines of evidence may interact in supporting a hypothesis.

Unfortunately, the Bayesian approach does not take us much further: it cannot, by itself, address the question of *conclusion uncertainty*, i.e. how the support for specific assumptions compounds and propagates onto a conclusion. As emphasised earlier, Bayesian networks rely on conditional probability tables/distributions that express the range of conditional probabilities of a node given the values its parents may assume (Koller & Friedman, 2009, 159). Furnishing these tables/distributions is difficult enough when deciding how different lines of evidence jointly bear on specific hypotheses (Das, 2004): there is no general theory that can tell us, say, how much confidence the combination of statistical and mechanistic evidence confers onto a causal conclusion beyond the sum of the confidence conferred by each

[11] I will subsequently refer to the approach offered by Landes et al. (2018) simply as 'the Bayesian approach'. This is not to suggest, however, that other approaches discussed, or indeed the approach I develop myself, are not also heavily Bayesian in character.

[12] See Wüthrich (2016) discussing analogous ideas pursued in the IPCC Assessment Report.

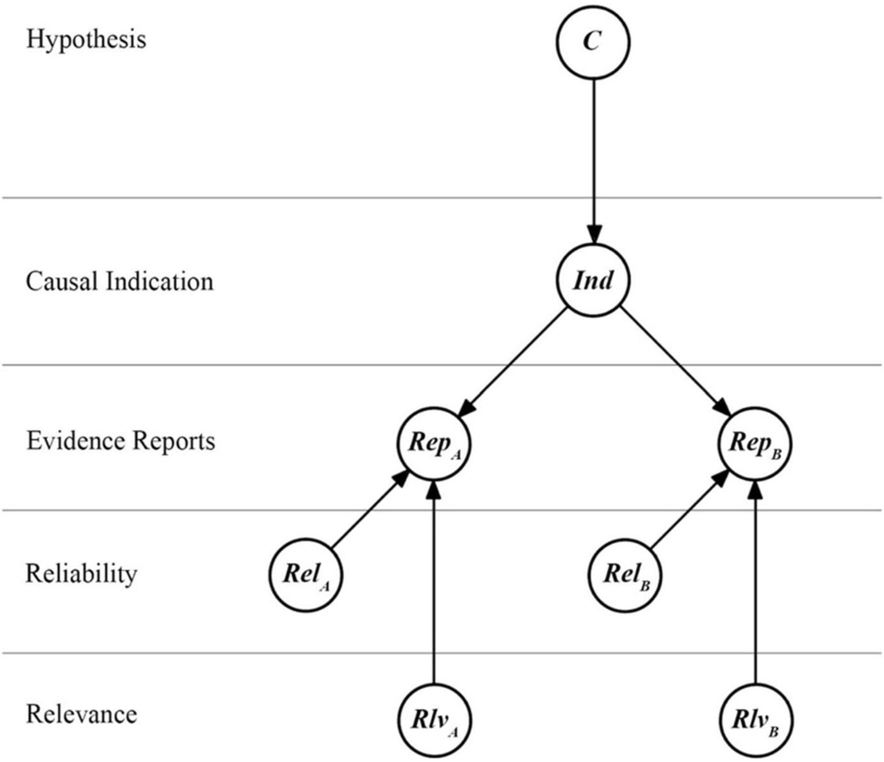

Fig. 1 Bayesian network after Landes et al. (2018)

type of evidence alone. We need to have a full-bodied, local theory of evidence to make such judgments (see Reiss, 2015).

Furnishing conditional probability tables/distributions is even more difficult when trying to tell how the support for different assumptions involved in a complex inference bears on a conclusion underwritten by these assumptions. As emphasised earlier, the way in which our confidence in a conclusion responds to the confidence we have in specific assumptions can be a complex function that depends on the causal structure of the systems under investigation and what roles our assumptions play in our inferences. In articulating how support for specific assumptions bears on our conclusion, we hence need to investigate not only how likely they are to be true, but also how the features that our assumptions pertain to hang together causally, and how our assumptions hang together inferentially, which must include considerations of evidential interdependence.[13] What is more, to appreciate the full extent of

[13] To be sure, most of the time the causal and inferential go hand in hand: an assumption pertaining to a feature that is causally more relevant to an effect will be more important inferentially precisely for that reason. But there can also be cases where the two can come apart, e.g. when an assumption is causally not especially important but inferentially highly relevant because supporting it also helps other, more important but difficult-to-support assumptions, e.g. people who wash their hands before eating probably wash their hands after going to the bathroom, and while the former might not be terribly relevant to whether they spread germs around the office, the latter is.

uncertainty that surrounds our conclusions we should also consider what would happen to our conclusions if our assumptions turned out false.

To appreciate these points, consider another example where two assumptions $A1$ and $A2$ are involved in an extrapolation. Suppose we have learnt in population A that a variable Z moderates how effective a social policy intervention is: the higher Z, the more pronounced the effect. In deriving a conclusion that our intervention will work in a novel target B, we hence need to assume ($A1$) that Z is suitably realized there. But this is not quite enough: we also need a further assumption, $A2$, which asserts that Z plays the same causal role in the mechanism in B as in A (cf. Cartwright, 2013a). Clearly, $A2$'s truth is a prerequisite for $A1$'s relevance – if Z does not play a role in the mechanism in B at all, or indeed the same role as in A, then we do not need to care about whether Z is suitably realized in B. So how does the support for $A1$bear on our conclusion? This will be a function of $A2$'s support, too, since the features to which they pertain hang together causally. More generally, we can say that encoding how evidence concerning some causal features (such as parameters, distributions of variables; in our case $A1$) bears on a conclusion via conditional probability tables/distributions may already presuppose knowledge of the very mechanisms that we might be uncertain about (i.e. strong support for $A2$). If $A2$ is strongly supported, the support for $A1$ matters a great deal for our conclusion; if $A2$ is unsupported (or indeed false), it does not matter at all. And if there is uncertainty surrounding $A2$, we may want to consider what happens to our effect of interest under each of the causal arrangements afforded by this uncertainty, and weighted by how likely each scenario is, given our evidence for $A2$.

To be sure, the Bayesian approach does not entirely ignore these issues. First, its proponents recognize the need to account for evidential interdependencies. If they did not, their approach would not be suited to integrating different kinds of evidence (e.g., statistical and mechanistic as per Russo & Williamson, 2007). Second, proponents of the Bayesian approach also claim that their account helps us manage problems of external validity and extrapolation (Landes et al., 2018, 24; see also Poellinger, 2020). Specifically, the relevance nodes figuring in their approach are intended to capture whether populations are sufficiently similar for evidence about A to be relevant to claims about B.

Despite promising ambitions, I do not think that the Bayesian approach can fully address the challenges looming in extrapolation. In practice, we cannot *assume* that analysts have a ready-made, high-dimensional conditional probability distribution/table that comprehensively encodes how varied evidence for different assumptions supports conclusions that depend in intricate ways on other assumptions we are simultaneously uncertain about. And while I do not take issue with *representing* judgments about how evidence from A speaks to queries about B by relevance nodes, this representation is only useful if the underlying problem of extrapolation is already solved: we must already know how evidence from A speaks to questions about B, but precisely this can be unclear in extrapolation contexts where severe uncertainty may persist about whether A and B are relevantly similar. Here, I consider cases where this cannot be assumed, where we face an ongoing extrapolation problem that involves significant uncertainties, and where producing a conclusion and acting on it requires articulating and managing these uncertainties. So, while the

Bayesian account can tell us how evidential support propagates 'upwards' towards specific assumptions, we need additional layers of analysis to consider how support propagates 'sideways', i.e. from a collection of causal assumptions onto a conclusion that depends on these assumptions in intricate ways. Let me outline a second approach that offers a useful rationale to help address this need.

3.3 The confidence approach

Roussos et al. (2021) address issues of uncertainty arising when climate scientists use ensemble models to predict rare climate events. Here, researchers are often faced with substantial uncertainty concerning whether the different models included in an ensemble adequately represent the earth's climate system (see e.g. Parker, 2013). Although Roussos et al.'s approach is not intended to address problems of extrapolation, the issues of uncertainty faced here are similar to those arising in model-based climate event forecasting: there exists deep uncertainty about whether important similarities between an experimental or model system and a target system of interest obtain.

The framework that Roussos et al. employ is a modified version of the *Confidence Approach* (Hill, 2013; Bradley, 2017), which seeks to capture how the weight of a body of evidence, as well as context-specific features (e.g. decision-making stakes, desired levels of confidence, attitudes towards uncertainty), bear on the confidence investigators may have in a model prediction: it provides a second-order assessment of how strongly a first-order prediction is supported by evidence. To provide such assessments, Roussos et al. use the idea of *nested intervals*, i.e. probability intervals of varying precision derived from different subsets of models from an ensemble. Their approach allows analysts to gauge how much confidence different probability intervals each enjoy, given how many models underwrite them and how competent these models are individually. Figure 2 illustrates.

Here, a more precise prediction interval, such as *A*, is only supported by a smaller set of model predictions (bottom three), so the confidence it might enjoy will be lower than for a wider interval (e.g. any top three interval). This approach hence articulates in an intuitive way how an agent's overall uncertainty hinges on the outputs of a range of models that differ in assumptions, and how confident the agent may be in specific predictions as a function of the support they enjoy over the range of existing uncertainty.

The Confidence Approach embodies a crucial rationale that figures prominently in a wide range of areas: to appreciate a state of uncertainty, we must consider what happens if things are different from what we initially assume. For instance, sensitivity and robustness analyses are a common tool in economics and other fields to explore whether estimates are robust under changes in the validity of identification assumptions (Manski, 1990, 2008; Rosenbaum, 2002). Similarly, cost-benefit analysis often relies on Monte-Carlo-based sensitivity analyses for telling whether the cost-effectiveness of interventions is robust over existing uncertainties (e.g. Whittington et al., 2020, 214). A similar rationale figures centrally in error statistics, where misspecification testing plays a crucial role in probing and validating statistical models (Mayo & Spanos, 2004), and in physics and engineering, where researchers use ensembles of computational physics models to quantify the uncertainty surrounding predictions of real-world engineering systems' behaviors (see Roy & Oberkampf, 2011). Surprisingly, despite its popularity, this rationale has not been thoroughly applied to problems of extrapolation, even though it seems highly suitable for exploring the uncertainties arising there.

4 The support graph approach

Our goal is now clearly in view: we need to tell how varied support for varied and potentially interdependent assumptions bears on the confidence in a causal conclusion. The Bayesian approach falls short of this goal, but delivers crucial resources for amalgamating varied evidence to speak to *specific* assumptions. The Confidence Approach recommends a useful rationale: we should consider how our predictions change if things are different from what we might initially assume (e.g. that two populations are similar). Let me outline an approach that systematizes and tailors these ideas to address the challenges arising in extrapolation: the *support graph approach* (SGA).

A clarification on aims upfront: my presentation of SGA aims at outlining a practical strategy that can help analysts explore and manage uncertainty in extrapolation. Oriented towards practice, SGA seeks to take some of the likely complications experienced by practitioners into account (e.g. limited epistemic and computational resources, the need to consider practical consequences of mistaken inferences, etc.) and outlines concrete avenues for analysts to tailor SGA to those needs and constraints. That said, SGA is also amenable to a more principled and normative interpretation, according to which SGA is a procedure that analysts *should* adopt to address issues of uncertainty. While I do not specifically defend such an interpretation here, I will emphasise several ways in which SGA facilitates normatively compelling goals, including: systematically exploring possible disparities; being sensitive to what causal and inferential roles our assumptions play; considering what assumptions are most in need of support, and so on. With SGA's aims clarified, let me outline its general structure before expanding on the concrete procedures it proposes to explore and manage uncertainty.

SGA combines three layers of analysis. The first is the *causal model layer*. Understanding which assumptions are required for an extrapolation is a thoroughly causal endeavour: it involves understanding the causal makeup of an experimental population, what similarities and differences between populations are important, and so on. These issues are best addressed by using causal models, as such models help us make explicit what assumptions are needed and what support they require (Khosrowi, 2021). For the purpose of outlining SGA, I rely loosely on the causal graph-based approach championed by Pearl and others (Bareinboim & Pearl, 2012, 2016; Pearl & Bareinboim, 2014), which draws on a combination of graphical causal models (called *directed acyclic graphs*, DAGs) and corresponding structural causal models (SCMs), and offers powerful analytical resources that permit sophisticated causal inferences. I will not discuss the details of this approach here – what is important

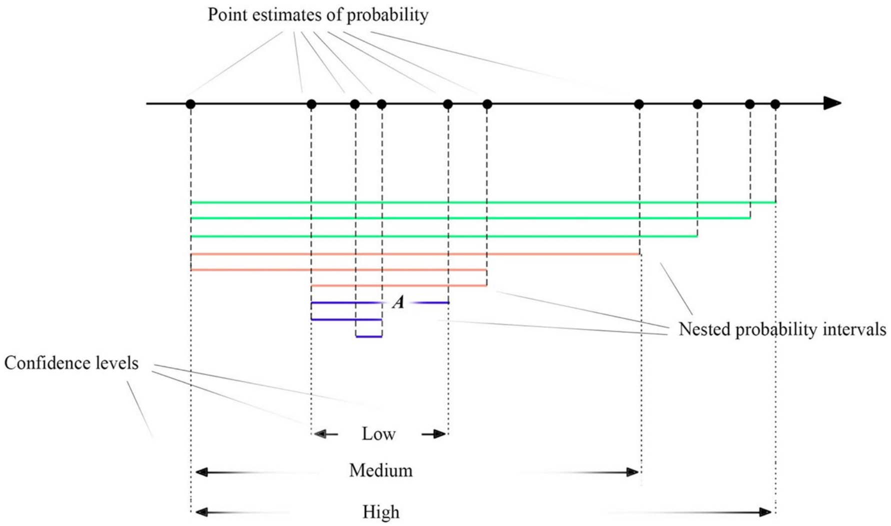

Fig. 2 The confidence approach after Roussos et al. (2021)

is that it can help answer a broad range of causal queries and tell what we need to assume for doing so.[14]

The second layer of SGA is the *support layer*. It focuses on *local* support: how well are our assumptions supported individually? As suggested earlier, this layer can draw extensively on the Bayesian approach, where the support for each assumption can be amalgamated using an inferential Bayesian network in the way that Landes et al. (2018) envision. Of course, important questions remain about what evidence to seek out, and how, exactly, we should integrate it. I will not engage with these issues here beyond pointing out that these empirical and theoretical challenges must be addressed.

The third layer, called the *relevance layer*, focuses on what role our assumptions play in our inference. It investigates how specific assumptions matter for a conclusion, how our conclusion varies with respect to whether they are satisfied, and brings together the *local* support for individual assumptions gathered at the support layer to bear on issues of *global* support: what confidence are we entitled to have in a conclusion, given the support our assumptions enjoy individually, their role and relevance in our inference, and taking into account their interdependencies. This layer draws on the rationale provided by the Confidence Approach.

With the general functions of these layers distinguished, let me elaborate how they can be integrated to facilitate structured assessments of uncertainty and

[14] The ability of Bareinboim and Pearl's framework to answer causal queries sets it apart from inferential uses of Bayesian networks, such as by Landes et al. (2018). SGA focuses on *causal* Bayesian networks for the causal model layer, and on inferential ones for the support layer.

confidence in extrapolation. SGA proceeds in three stages[15]: the first is concerned with structural uncertainties, the second with quantitative uncertainties, the third refines the results.

4.1 Stage 1: Mapping out structural uncertainty

SGA's main aim in tackling structural uncertainty is to identify a set of possible causal models $\boldsymbol{M}^*$ which encodes how a target population may differ from an experimental population at the level of the structure of causal mechanisms. This inquiry begins with a reference model $M_A$that captures what is known about the experimental population and proceeds by generating new, alternative models that depart from M_A with respect to specific features we are uncertain about.

How does SGA individuate these alternative models? A causal DAG-based approach following Bareinboim and Pearl permits two ways of doing so. One focuses on the graphical causal model itself and the causal relationships R encoded by causal arrows, e.g. $X \rightarrow Y$. Here, we might be uncertain about whether a relationship R between X and Y obtains in a target and hence individuate two possible models, one with and one without R. A second way of individuating models focuses on the structural causal models (SCMs) that correspond with our graphical models. Here, causal relationships are encoded by structural equations, and models are individuated by writing down different structural equations, e.g. $Y = f_1(X, Z) = X + Z$ to indicate that both X and Z additively contribute to Y, or $Y = f_2(Z) = Z$ to indicate that only Z causes Y.

There are cases where the latter approach seems more appropriate: in addition to encoding qualitative information that is also contained in a graphical causal model, e.g. *whether* X and Y are related at all, SCMs can encode richer parametric information about *how* X and Y are related. Consequently, this allows us to express finer-grained uncertainties about these relationships. For instance, we might face *parametric* uncertainty about how Y depends on X and Z, e.g. in an additively separable way, $Y = f_1(X, Z) = X + Z$, or rather in an interactive way, $Y = f_3(X, Z) = X * Z$. Graphical causal models do not offer sufficient resolution to capture the difference between these cases, representing both by the same causal DAG.[16]

To keep the presentation of SGA simple, I focus here on the first, graphical way to individuate models. In doing so, my focus is on qualitative issues of whether or not there are causal relationships R between pairs of variables, noting that each qualitative difference between graphical models would also be accompanied by a

[15] It may seem confusing that SGA involves three layers as well as three stages. To clarify, the three layers indicate the kinds of analysis that SGA seeks to integrate, which existing approaches have not managed to do. The stages discussed here describe the concrete procedure by which SGA recommends we should articulate, explore and manage uncertainty. The stages, in turn, each embody some or all of the layers of analysis that SGA integrates.

[16] This is not to say that graphical models do not allow for causal interaction or non-linear relationships. In the spirit of Pearl's framework, the graphical models discussed here should be understood as based on a nonparametric SCM, which allows that any structural equation may be non-linear and/or involve interactions.

corresponding difference at the level of SCMs. In principle, however, SGA can draw on the full resources of Bareinboim and Pearl's causal DAG-based framework, and is hence prepared to individuate models at the SCM level in addition to, or instead of, at the more coarse-grained DAG level. Importantly, taking this route would not change how SGA proceeds; it simply allows us to encode more possible disparities, and hence fine-grain $\boldsymbol{M}^{*}$ to account for parametric uncertainties as well.[17]

With this in mind, let me explain how $\boldsymbol{M}^{*}$ is constructed, starting from a reference model M_A. Assuming mechanisms are modular (see e.g. Steel, 2009), each of the causal relationships R comprising $M_A$may be present or absent in a target.[18] Figure 3 illustrates a model M_A that captures the microfinance example from earlier.

Drawing on the Bayesian approach, we can now bring evidence to bear on the assumptions encoded in M_A. For each causal relationship R in M_A, such as the relationship $X \rightarrow I$, a measure of support/confidence S can be computed to encode how confident we are that the relationship holds in the target.[19] In a principled analysis of uncertainty, SGA would demand that analysts consider *all* possible disparities that cannot be completely ruled out by available evidence. Here, anything less than full confidence that R is instantiated in a target would press analysts to generate an alternative model not including R to later consider. However, SGA is also prepared to relax these demands: on a more pragmatic interpretation, a reasonable aim might be to distinguish broadly between those relationships/features we are sufficiently confident in and those we are too uncertain about[20]. To make this distinction, a threshold of confidence $\alpha \in [0,1]$ can be used. If our confidence S in R exceeds α, say 0.9, then R is assumed to hold and no alternative model is generated. If it fails to exceed α, the uncertainty surrounding R will subsequently be explored.

Say we are uncertain about the relationship $R_1 : X \rightarrow I$; our confidence S that it holds in B is 0.7. To capture this uncertainty, we generate an additional model, M_{B1}, that does not contain R_1but is otherwise identical to M_A. Both models will subsequently be used to explore how our predictions hinge on the uncertainty surrounding R_1. To help them bear on this assessment, each of the two models, M_A and M_{B1}, is assigned a weight w_n that corresponds to the confidence in either model as a function of our confidence in R_1. In our present case, the weight for $M_A$would be

[17] I thank an anonymous reviewer for pressing me to say more on these issues.

[18] By 'modular' I mean the assumption that causal relationships comprising a mechanism can, in principle, be changed without thereby altering other relationships. In the present DAG-focused context, I also take this to mean that not only structural equations but also causal arrows can be changed in this way. These assumptions are not always plausible. When they fail, analysts should consider how this bears on the success of their inferences and interventions (if possible). Importantly, however, SGA can still proceed as envisioned, but must account for any non-modular dependencies among causal relationships. For instance, if the absence of R1 implies that another relationship R2 changes to R2* (or does not obtain), then all models that do not contain R1 should exhibit these features too (other things being equal).

[19] As anticipated earlier, we might alternatively focus on structural equations here and assess how confident we are that a specific equation correctly describes the target.

[20] For instance, α may be set to lower values to reflect that the stakes of being wrong about a causal feature are low, or that there is decreasing marginal epistemic utility as the confidence in specific assumptions approaches unity.

$w_1 = S = 0.7$and for M_{B1} it would be $w_2 = (1 - S) = 0.3$.[21] The weights reflect that our evidence allows us to lean towards $M_A$but that M_{B1} needs to be considered, too, since we cannot rule out that this is what the target looks like.

To capture further structural uncertainties, additional models M_{B2}, M_{B3}, M_{B4}, etc. may be added to $\boldsymbol{M}^*$ in the same way for each relationship/feature we are uncertain about. Figure 4 illustrates.

In addition to considering how relationships encoded in M_A might fail to obtain in a target, analysts may also wish to consider extraneous causal relationships in a target that are not present in M_A, e.g. an alternative causal pathway from X to Y. Of course, there may be an intractably large number of extraneous relationships we could stipulate. To keep the cardinality of $\boldsymbol{M}^*$ manageable, a useful relaxation of SGA for practice is to only consider relationships that are consistent with, or indicated by, background knowledge.

The result of this first stage is a set of models $\boldsymbol{M}^*$ that captures possible causal arrangements in the target as afforded by our structural uncertainties. Naturally, the confidence threshold α has substantial bearing on how many models will enter $\boldsymbol{M}^*$, and the choice of α is non-trivial: how much support is enough to decide that we can lean on an assumption as fixed, and disregard the remaining uncertainty surrounding it? Nor is it clear that a single threshold is always appropriate, as analysts may wish to reject or add specific models to $\boldsymbol{M}^*$ based on additional considerations, such as the non-epistemic stakes involved in mistakenly ignoring a specific disparity. α is hence only a conceptual placeholder, and analysts can and must exercise discretion concerning which models to consider.

4.2 Stage 2: Exploring structural and quantitative uncertainty

With $\boldsymbol{M}^*$ in place, we can now explore both structural and quantitative uncertainties in a unified and systematic way. Our structural uncertainties are already captured by the models in $\boldsymbol{M}^*$. Quantitative uncertainties concern the values of structural parameters and the values/distributions of variables that figure in the models $\boldsymbol{M}^*$. While uncertainties regarding observable variables are often less acute (unless they concern difficult-to-measure variables), uncertainties concerning structural parameters (e.g. the rate at which individuals convert investments into income) are common.

Quantitative uncertainties can be explored in the following way: for each of the models in $\boldsymbol{M}^*$, we can derive a set of predictions of the effect of interest, i.e. $P(Y|\text{do}(\text{x}))$ in the language of the causal DAG-based approach, over existing quantitative uncertainties.[22] Consider our model M_A again and suppose we are uncertain about the value of a structural parameter β that shapes the $I \rightarrow H$ relationship. Suppose we know that β may be in the range $[a, b]$. To explore the consequences of

[21] When further partitioning the model space to account for additional uncertainties, model weights from all steps are multiplied for each model, which ensures that model weights sum to unity.

[22] For simplicity, I consider the intervention variable to be dichotomous, i.e. $X \in \{0,1\}$. Moreover, following Bareinboim and Pearl's conventions, I characterize causal effects as outcome distributions conditional on an intervention that sets $X = x$, i.e. $P(Y|\text{do}(\text{X} = \text{x}))$. By contrast, large parts

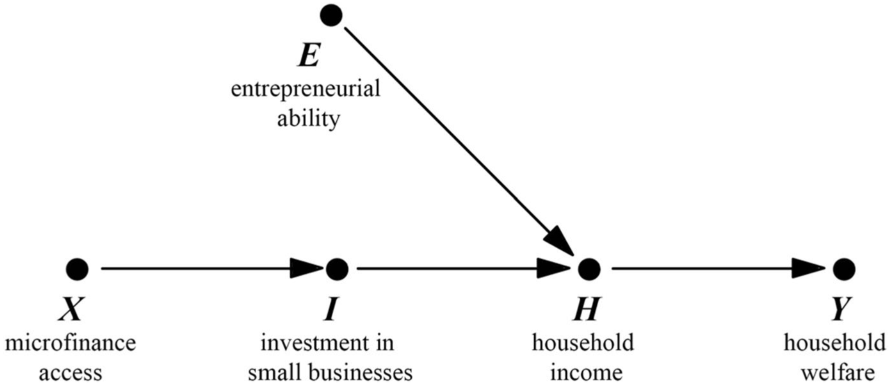

Fig. 3 Causal model M_A of the microfinance effect

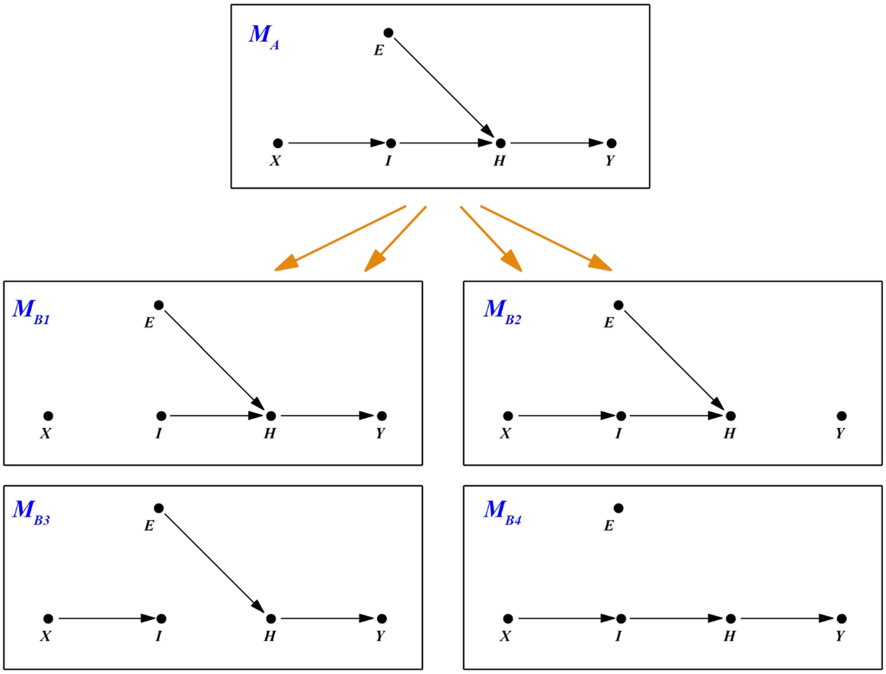

Fig. 4 A set of possible models $\boldsymbol{M}^*$

Footnote 22 (continued)
of the empirical social science literature understand effects rather as *differences* in the expectation between an outcome under different values of an intervention variable, e.g. as average causal effects $ACE(X) = E(Y|\text{do}(X = 1)) - E(Y|\text{do}(X = 0))$. In principle, SGA can proceed in terms of such effect measures as well, but since doing so would require that the figures used for illustration here display two distributions/means at a time for each effect, I choose to work with Bareinboim and Pearl's (graphically) leaner conception.

this uncertainty, we can furnish two or more predictions across the range of values afforded by it. In the simplest case, we may consider only the endpoints $\beta = a$ and $\beta = b$. More comprehensively, we may furnish predictions across finer-grained partitions of our uncertainty surrounding β. In the standard fashion of sensitivity analysis, this allows us to tell how much an effect changes over variation in β afforded by our uncertainty. Figure 5 illustrates.

To provide a comprehensive assessment of structural and quantitative uncertainty, this process is repeated for *all* quantitative uncertainties regarding other parameters or the values/distributions of variables, and over *all* models in $\boldsymbol{M}^*$. The result of this process is a family $\boldsymbol{P}^*$ of probability distributions $P(Y|\text{do(x)})$ that traces out the whole extent of quantitative uncertainty over the whole extent of structural uncertainty captured by $\boldsymbol{M}^*$.

In furnishing an assessment of our overall state of uncertainty, the final step is to integrate the distributions comprising $\boldsymbol{P}^*$ into an overall outcome distribution. This integration considers the model weights w_n assigned earlier. Each prediction weighs only as heavily as the joint support in its favor.[23] As above, analysts may exercise additional discretion when deciding how to integrate predictions; e.g. they may use a higher-order weighting function to emphasise specific aspects of the predictions, e.g. weigthing the tails of a distribution more heavily to reflect risk-averse preferences. Figure 6 illustrates different probability distributions $P(Y|\text{do(x)})$ from different models combined into a single outcome distribution using model weights w_n.

Together, the first two stages provide analysts with a comprehensive assessment of an overall state of uncertainty that reflects the structural and quantitative uncertainties they experience, and accounts for the relative support for specific assumptions.

4.3 Stage 3: Refinement

The final, third stage of SGA seeks to refine the insights obtained. Given information about an overall state of uncertainty, analysts may wish to learn how acquiring additional evidence can help reduce this uncertainty. Here, analysts can proceed by fixing specific structural and quantitative degrees of freedom at stages 1 and 2, e.g. assuming that β is known with precision, or that R_1 is known to hold in a target, and explore how doing so changes an overall state of uncertainty. This provides an intuitive grasp of how relevant specific features and assumptions are: other things being equal, fixing an important feature will reduce the overall uncertainty more dramatically than fixing an unimportant feature. Such investigations can guide analysts in identifying assumptions that are highly relevant but poorly supported or constrained, thus capturing the key intuition behind Cartwright and Stegenga's (2011) WL principle. Detailing this intuition, SGA's finer-grained resources can help analysts identify specific combinations where the cost of acquiring additional evidence and its

[23] Similar to Roussos et al. (2021, 32), I will not say more on what integration methods should be used for this purpose. Many options are available, including various kinds of Bayesian model averaging techniques (see Leamer, 1978; Friedman & Koller, 2003), and I assume that analysts make informed choices about which techniques suit their purposes.

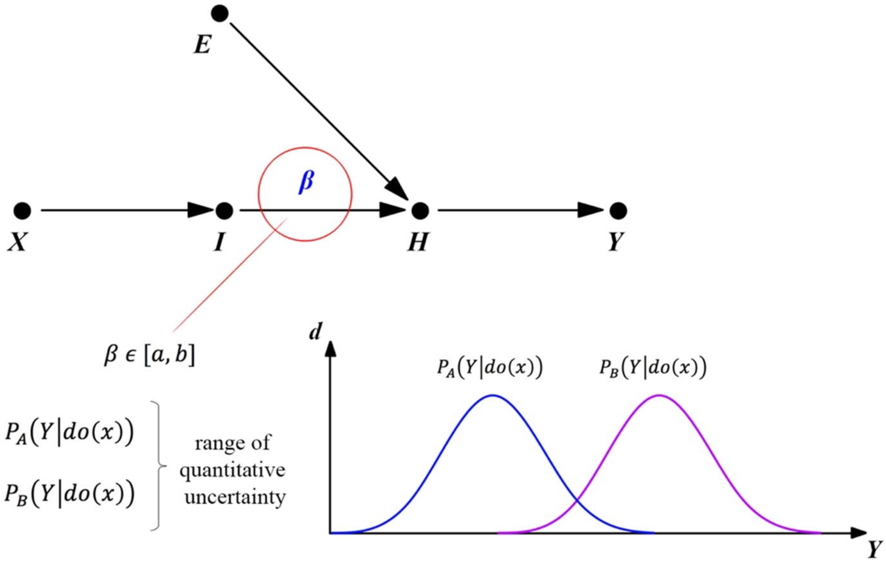

Fig. 5 Exploring quantitative uncertainty

effectiveness in reducing uncertainty strike a good balance. Stage 3 then consists in iterating this process until a desired level of confidence is reached or available epistemic resources are exhausted. With this sketch of SGA in place, let me discuss its broader virtues and allay some important worries.

5 Virtues and vices

A central virtue of SGA is that it establishes productive, symbiotic relationships with a range of existing frameworks. First, SGA draws on a rich set of capabilities afforded by the causal DAG-based framework. There are, of course, ways to achieve successful extrapolation without involving causal models, but many authors agree that models can perform crucial functions in facilitating better inferences (Pearl, 2009; Cartwright & Stegenga, 2011; Cartwright, 2013a; Khosrowi, 2021). By putting models at the heart of extrapolation, SGA helps analyst draw on these resources, and consider how assumptions work together in enabling a conclusion. SGA does not only harness useful capabilities of this framework, it also builds a symbiotic relationship with it: one important criticism of the causal DAG-based approach is that it fails to provide recipes for supporting the substantive assumptions it involves and expressing the uncertainties surrounding them (Hyttinen et al., 2015). SGA complements Bareinboim and Pearl's approach by addressing precisely these issues, utilizing its resources to furnish systematic explorations of structural uncertainty and tracing out how such uncertainty compounds with regard to the ultimate conclusions derived about a target – something that their approach cannot provide on its own. Second, SGA accommodates and extends crucial capabilities offered by

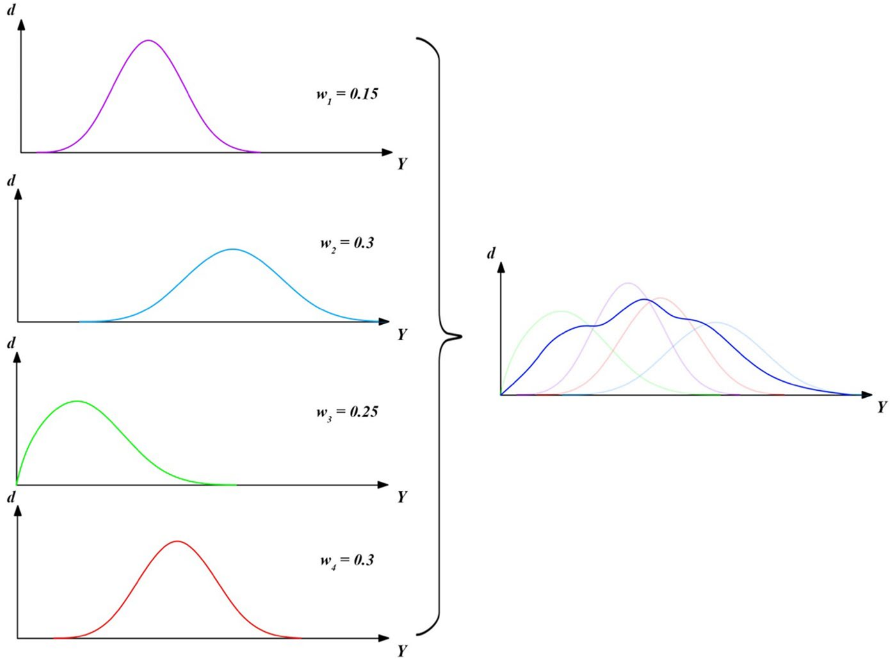

Fig. 6 Integrating predictions into final outcome distribution

the Bayesian approach, using its resources to let evidence speak to specific assumptions, while helping the approach bear on larger questions of conclusion uncertainty it cannot address by itself. Third, SGA draws on the central rationale offered by the Confidence Approach: to characterize a state of uncertainty, we need to consider how our conclusion may vary if things differ from what we initially assume. SGA should hence be an attractive approach for investigators already familiar with analyses that draw on this rationale. Finally, SGA inherits and refines intuitions provided by Cartwright and Stegenga's (2011) WL principle: it maintains that focusing on poorly supported and highly relevant assumptions is crucial, but also allows us to systematically explore an overall state of uncertainty, determine what it hinges on, and extract insights about how to manage it.

Another important virtue of SGA is its comprehensive character: if followed thoroughly (as we may demand on a normative interpretation), it makes it unlikely that we miss important possibilities for how a target might differ from an experimental population by pressing analysts to perform a comprehensive search for relevant differences. Since this search does not have to rely on antecedent guidance concerning what scenarios to consider (though it may do so), it can, in principle, be automated. Such an approach can be advantageous as it minimizes the scope for wishful thinking (e.g. focusing on evidence to confirm hoped-for similarities), and can help overcome important blind spots. Considering what we know about the causal makeup of an experimental population is of course a useful starting point to guide investigation – but serious problems of extrapolation are about cases where differences are likely but difficult to anticipate, so considering a fuller range of possible disparities is important for appreciating what uncertainties we face.

5.1 Overdemandingness

On the heels of emphasising the comprehensive ambitions of SGA, however, comes an important worry: that it is very demanding, and perhaps overly so. Let me discuss and respond to several variants of this concern and highlight further features and virtues of the approach[24].

First, how do analysts get access to good causal models of the experimental population? Without such a baseline, it seems difficult to imagine how they could systematically explore relevant disparities, let alone in an automated way. This is an important concern. Using models to assist with extrapolation is becoming more widespread, such as in the realist evaluation literature (Pawson, 2006, 2013; Astbury & Leeuw, 2010), which puts emphasis on modelling the processes and mechanisms by which an intervention is supposed to work. But even when models are used, important concerns remain about their quality. One might wonder what good models are if they amount to little more than a *vision* of how an intervention is hoped to be effective, rather than an empirically grounded account of what mechanisms actually govern its effectiveness. Mechanistic knowledge is hard to come by and we should not expect it to be available whenever it is needed.

My response here is to emphasise that insisting on the use of models, even if they are (initially) bad, can still be epistemically beneficial (see Khosrowi, 2021). As highlighted earlier, a model-based approach to extrapolation has the important advantage of pressing investigators to make their causal reasoning explicit. *A* model, rather than no model, is often a useful first step towards a good model, and even models that get things wrong can be useful, as long as we consider how things might differ from what we initially assume – and this is precisely what SGA can help us with.

A second variant of the overdemandingness worry piggybacks on the first: SGA is intractable. Even if we had all the causal knowledge to get things started, the process of building $\boldsymbol{M}^*$, weighing specific assumptions and whole models according to their support, furnishing predictions over various degrees of freedom, and assessing the relevance of assumptions across the model space, is simply too complicated to be useful in practice.

I have two replies to this worry. The first is that SGA can be scaled in various dimensions. We can scale the number of models going into $\boldsymbol{M}^*$ by imposing different thresholds α, and model generation can be additionally policed by criteria of plausibility and salience – we don't need to consider that income might cause age but we might want to consider whether race plays different causal roles in different populations. We can also scale the approach regarding its detail: assessing how well-supported specific assumptions or whole models are does not need to

[24] I thank Jaakko Kuorikoski for raising these concerns.

proceed in a quantitative way, but can be done qualitatively and with substantial scope for rough-and-ready judgment. Finally, assessments of overall uncertainty can be expressed qualitatively, too, tallying up weighted reasons to think that an effect might differ qualitatively from a reference prediction, and adding qualitative judgments of support and relevance to map out an overall state of uncertainty in broader strokes. This will, of course, compromise the precision of the uncertainty assessments obtained, but that may be an acceptable price to pay depending on the stakes and constraints at hand. SGA hence offers various meaningful decision points where contextual information can help shape the details of the inquiry. My second reply is about the bigger picture: on a normative reading, SGA can be understood to characterize a guiding ideal. It is a general template for a sound procedure to assess and manage uncertainty, but not an all-or-nothing affair. As with many ideals, we can realize some of its procedural virtues, approximate it in specific respects, and reap important benefits without following its demands exhaustively and exhaustingly. Doing better rather than best still beats doing poorly.

A third variant of the overdemandingness worry concerns cost-effectiveness: if we follow SGA thoroughly, the costs of analysis may outweigh the costs of remaining ignorant. There are many cases where this concern has bite – even if SGA is not entirely overdemanding, it is not obvious that its benefits are significant enough to make it attractive to practitioners. This worry is only partially addressed by emphasising that SGA can be scaled variously. Here, I want to emphasise instead that the scope of SGA is not universal: not all extrapolation problems benefit from the sorts of assessments that SGA helps furnish, e.g. because they do not involve the kinds of deep, structural and quantitative uncertainties that motivate the approach, or the costs of wrong predictions are not significant. So, SGA should be understood to focus on those cases where the stakes are high, the problems to be tackled complex, and analysts are prepared (and justified) to invest substantial resources in analysis, e.g. large-scale policy decisions involving significant costs and considerable uncertainty.

Finally, a fourth variant of the overdemandingness worry focuses on the practical utility and disutility of learning about uncertainty. In short, we might worry that recognizing an overall state of uncertainty that exceeds what decision-makers anticipated might paralyze them, suggesting it might sometimes be better to remain ignorant of it. Here, it seems important to emphasise that such uncertainties exist independently of SGA – the approach just reveals them. My stance here follows a common ideal: we should make ourselves aware of the extent of our ignorance and uncertainty and SGA can help promote this ideal. In adopting this stance, SGA once more discloses some normative aspirations, at least conditional ones. To the extent that analysts care about articulating and managing uncertainty, SGA specifies the sorts of rationales they should follow, to an extent that is feasible given the pragmatic constraints of the context. Deviations from the ideal are acceptable, but, if taken too far, may come on pain of making poor decisions based on inadequate assessments of uncertainty. SGA, I contend, provides a promising approach for those interested in doing better.

6 Conclusion

Extrapolating causal effects from experiments often involves difficult-to-support assumptions surrounded by significant uncertainty. Existing approaches for articulating uncertainty do not, by themselves, provide assessments that take the complex nature of extrapolative inference into account. The Bayesian approach (Landes et al., 2018) is useful for articulating local uncertainties, but cannot provide global assessments. The Confidence Approach (Roussos et al., 2021) offers a compelling rationale for exploring uncertainty, but needs to be concretized to speak to issues arising in extrapolation. These are not shortcomings of existing approaches per se, but they suggest that additional layers of analysis are needed to meet the epistemic and practical needs of analysts and decision-makers, who, in the face of real stakes, will often need to know not only what concrete effects they may expect in a novel population, but also how confident they can be in these expectations.

I have characterized the *support-graph approach* (SGA) as a first sketch of an approach that helps address these issues. The virtues highlighted here suggest it is a worthwhile approach to explore further, and the concerns considered point to concrete avenues for refining it. Additional work is needed to spell out its details, including how to ensure SGA remains tractable at scale without compromising its ability to highlight important uncertainties. Subsequent steps in further detailing SGA may include a thorough formal rendition and implementing it computationally to help explore its concrete abilities in providing analysts with useful assessments of uncertainty and avenues to manage it.

Rather than providing a fully worked out formal account, the focus of this paper has been to articulate the limitations of existing approaches, to improve our understanding of how uncertainty in extrapolation may be addressed, and to characterize, in broad strokes, a promising approach for doing so. So, while SGA is perhaps not quite ready to be implemented in practice, we might be content for now with how it improves over existing approaches by systematizing a crucial rationale: to tell how confident we can be in an extrapolation, we need to consider what happens to our predictions across the full range of uncertainty we experience. SGA makes important progress on operationalizing this rationale, helping analysts explore the sources and consequences of uncertainty and identify ways for mitigating it. In virtue of this, SGA can be a useful starting point for helping experimental evidence speak more confidently to questions that decision-makers face.

Acknowledgements I wish to thank Wendy Parker, Julian Reiss, Nancy Cartwright, Katherine Puddifoot, Robert Northcott, Mathias Frisch, William Peden, Joe Roussos and members of the CHESS and K4U research groups at Durham University for valuable discussions that helped shape the ideas developed in this paper. Moreover, I would like to thank the audiences at BSPS 2021 and EPSA 2021 for raising several important points that helped improve my arguments. Finally, I would like to thank two anonymous reviewers for their many very helpful comments, criticisms and suggestions, which helped significantly improve the ideas developed in this paper.

Funding Open Access funding enabled and organized by Projekt DEAL. My work on this paper was financially supported by an AHRC Northern Bridge Doctoral Studentship (grant number: AH/L503927/1), a Durham Doctoral Studentship, and a Royal Institute of Philosophy Jacobsen Studentship.

Declarations

Competing interests No additional financial or non-financial interests are related to the work submitted.

References

Astbury, B., & Leeuw, F. (2010). Unpacking black boxes: Mechanisms and theory building in evaluation. *American Journal of Evaluation, 31*(3), 363–381.

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives, 31*(2), 3–32.

Bareinboim, E., & Pearl, J. (2012). Transportability of causal effects: Completeness results. In Proceedings of the Twenty-Sixth Conference on Artificial Intelligence (AAAI-12), Menlo Park, CA.

Bareinboim, E., & Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, *113*, 7345-52.

Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology*. Oxford University Press.

Bradley, R. (2017). *Decision theory with a human face*. Cambridge University Press.

Cartwright, N. D. (2011). Predicting 'It will work for us': (Way) beyond statistics. In F. R. Phyllis McKay Illari, & J. Williamson (Ed.), *Causality in the Sciences*. Oxford Scholarship Online.

Cartwright, N. D. (2013a). Knowing what we are talking about: Why evidence doesn't always travel. *Evidence & Policy, 9*(1), 97–112.

Cartwright, N. D. (2013b). Evidence, argument and prediction. In *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*. The European Philosophy of Science Association Proceedings 2. Springer.

Cartwright, N. D., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford University Press.

Cartwright, N. D., & Stegenga, J. (2011). A theory of evidence for evidence-based policy. In P. Dawid, W. Twining., & M. Vasilaki (Eds.), *Evidence, Inference and Enquiry* (Proceedings of the British Academy). Oxford University Press.

Cowen, N., & Cartwright, N. D. (2019) Street-level theories of change: Adapting the medical model of evidence-based practice for policing. In N. Fielding, K. Bullock, & S. Holdaway (Eds.), *Critical reflections on evidence-based policing. Routledge frontiers of criminal justice* (pp. 52–71). Routledge.

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2008). Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics, 90*(3), 389–405.

Das, B. (2004). Generating conditional probabilities for Bayesian networks: Easing the knowledge acquisition problem. CoRR cs.AI/0411034.

Deeks, J. J., Higgins, J. P. T., & Altman, D. G. (2022). Chapter 10: Analysing data and undertaking meta-analyses. In Higgins, J. P. T., J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch (Eds.) *Cochrane Handbook for Systematic Reviews of Interventions version 6.3*. www.training.cochrane.org/handbook. Accessed 28 Feb 2022.

Duflo, E. (2018). *Machinistas meet randomistas: Useful ML tools for empirical researchers*. Summer Institute Master Lectures. National Bureau of Economic Research.

Friedman, N., & Koller, D. (2003). Being bayesian about network structure. A bayesian approach to structure discovery in bayesian networks. *Machine learning, 50*(1–2), 95–125.

Good, I. J. (1985). Weight of evidence: A brief survey. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics 2* (pp. 249–270). Elsevier.

Hill, B. (2013). Confidence and decision. *Games and Economic Behaviour, 82*, 675–692.

Hotz, V. J., Imbens, G. W., & Mortimer, J. H. (2005). Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics, 125*, 241–270.

Hyttinen, A., Eberhardt, F., & Järvisalo, M. (2015). Do-calculus when the true graph is unknown. *In Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence (UAI'15)*. AUAI Press, Arlington, Virginia, USA, pp. 395–404.

Keynes, J. M. (1921). *A treatise on probability*. Macmillan.

Khosrowi, D. (2019). Extrapolation of causal effects – hopes, assumptions, and the extrapolator's circle. *Journal of Economic Methodology, 26*(1), 45–58.

Khosrowi, D. (2021). When experiments need models. *Philosophy of the Social Sciences, 51*(4), 400–424.

Khosrowi, D. (2022). Evidence-based policy. In J. Reiss & C. Heilmann (Eds.), *The Routledge Handbook of Philosophy of Economics* (pp. 370–84). Routledge.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principle and techniques*. MIT Press.

Landes, J., Osimani, B., & Poellinger, R. (2018). Epistemology of causal inference in pharmacology. *European Journal for Philosophy of Science, 8*(1), 3–49.

Leamer, E. E. (1978). *Specification searches*. Wiley.

Manski, C. (1990). Nonparametric bounds on treatment effects. *American Economic Review, 80*, 319–323.

Manski, C. (2008). *Identification for prediction and decision*. Princeton University Press.

Mayo, D. G., & Spanos, A. (2004). Methodology in practice: Statistical misspecification testing. *Philosophy of Science, 71*(5), 1007–1025.

Muller, S. M. (2014). Randomised trials for policy: A review of the external validity of treatment effects. Southern Africa Labour and Development Research Unit Working Paper 127, University of Cape Town.

Muller, S. M. (2015). Interaction and external validity: Obstacles to the policy relevance of randomized evaluations. *World Bank Economic Review , 29*(1), 217–225.

Parker, W. (2013). Ensemble modeling, uncertainty and robust predictions. *Wiley Interdisciplinary Reviews: Climate Change, 4*(3), 213–223.

Pawson, R. (2006). *Evidence-based policy: A realist perspective*. SAGE.

Pawson, R. (2013). *The science of evaluation: A realist manifesto*. SAGE Publications.

Pearl, J. (2009). *Causality: Models, reasoning, and inference*. 2nd edition. Cambridge University Press.

Pearl, J., & Bareinboim, E. (2014). External validity: From Do-Calculus to transportability across populations. *Statistical Science, 29*(4), 579–595.

Peirce, C. S. (1878). The probability of induction. *Popular Science Monthly, 12*, 704–718.

Poellinger, R. (2020). Analogy-based inference patterns in Pharmacological Research. In B. Osimani (Ed.), *Uncertainty in Pharmacology* (338 vol.). Boston Studies in the Philosophy and History of Science. https://doi.org/10.1007/978-3-030-29179-2_5

Reiss, J. (2015). A pragmatist theory of evidence. *Philosophy of Science, 82*(3), 341–362.

Reiss, J. (2019). Against external validity. *Synthese, 196*(8), 3103–3121.

Rosenbaum, P. (2002). *Observational Studies*. 2nd edition. Springer.

Roussos, J., Bradley, R., & Frigg, R. (2021). Making confident decisions with model ensembles. *Philosophy of Science, 88*(3), 439–460. https://doi.org/10.1086/712818

Roy, C., & Oberkampf, W. L. (2011). A complete framework for verification, validation, and uncertainty quantification in scientific computing. *Computer Methods in Applied Mechanics and Engineering, 200*, 2131–2144.

Russo, F., & Williamson, J. (2007). Interpreting causality in the health sciences. *International Studies in the Philosophy of Science, 21*, 157–170.

Sidebottom, A., Tompson, L., Thornton, A., Bullock, K., Tilley, N., Bowers, K., & Johnson, S. D. (2018). Gating alleys to reduce crime: A meta-analysis and realist synthesis. *Justice Quarterly, 35*(1), 55–86.

Steel, D. (2009). *Across the boundaries: Extrapolation in biology and social science*. Oxford University Press.

van Eersel, G. G., Koppenol-Gonzalez, G. V., & Reiss, J. (2019). Extrapolation of experimental results through analogical reasoning from latent classes. *Philosophy of Science, 86*(2), 219–235.

Vivalt, E. (2020). How much can we generalize from impact evaluations? *Journal of the European Economics Association, 18*(6), 3045–3089.

Weed, D. L. (2005). Weight of evidence: A review of Concept and methods. *Risk Analysis, 25*(6), 1545–1557.

Whittington, D., Radin, M., & Jeuland, M. (2020). Evidence-based policy analysis? The strange case of the randomized controlled trials of community-led total sanitation. *Oxford Review of Economic Policy, 36*(1), 191–221. https://doi.org/10.1093/oxrep/grz029

Williamson, J. (2020). The feasibility and malleability of EBM+. *Theoria, 36*(2), 191–209. https://doi.org/10.1387/theoria.21244

Wüthrich, N. (2016). Conceptualizing uncertainty: An assessment of the uncertainty framework of the intergovernmental panel on climate change. In *EPSA15 Selected Papers: The 5th conference of the European Philosophy of Science Association in Düsseldorf.*