Generalization Bias in Science

[This is a penultimate draft of a paper forthcoming in *Cognitive Science*.
Comments very welcome.]

Uwe Peters[1]
(a) Leverhulme Centre for the Future of Intelligence, University of Cambridge, UK
(b) Center for Science and Thought, University of Bonn, Germany

Alexander Krauss
(a) CPNSS, London School of Economics, UK
(b) Spanish National Research Council, Barcelona, Spain

Oliver Braganza
(a) Institute for Experimental Epileptology and Cognition Research, University of Bonn
(b) Center for Science and Thought, University of Bonn

Abstract

Many scientists routinely generalize from study samples to larger populations. It is commonly assumed that this cognitive process of scientific induction is a voluntary inference in which researchers assess the generalizability of their data and then draw conclusions accordingly. Here we challenge this view and argue for a novel account. The account describes scientific induction as involving by default a generalization bias that operates automatically and frequently leads researchers to unintentionally generalize their findings without sufficient evidence. The result is unwarranted, overgeneralized conclusions. We support this account of scientific induction by integrating a range of disparate findings from across the cognitive sciences that have until now not been connected to research on the nature of scientific induction. The view that scientific induction involves by default a generalization bias calls for a revision of our current thinking about scientific induction and highlights an overlooked cause of the replication crisis in the sciences. Commonly proposed interventions to tackle scientific overgeneralizations that may feed into this crisis need to be supplemented with cognitive debiasing strategies to most effectively improve science.

Keywords: scientific induction; overgeneralization; bounded cognition; generalization bias; replication crisis

[1] Corresponding and main author: up228@cam.ac.uk

1. Introduction

We humans frequently encounter new situations in which we need to generalize from previous experiences to be able to respond adaptively (Wu et al., 2018). In doing so, we commonly use "induction", the cognitive process of inferring that "what is true of certain individuals of a class is true of the whole class, or that what is true at certain times will be true in similar circumstances at all [or most] times" (Mill, 1872/1974, p. 188). Induction is ubiquitous in human cognition[2] already early in life (Fisher et al., 2015). It can take many different forms (Hayes & Heit, 2018) and is often used instinctively (e.g., when children avoid a hot stove after a single burn; Leslie, 2017).

Induction is also fundamental in much of scientific reasoning (Bunge, 1960; Norton, 2005; Achinstein, 2010). We will focus here on reasoning in the cognitive, behavioral, and social sciences.[3] Specifically, the focus will be on generalizations that occur when scientists collect data from their study samples and extrapolate from them to broader populations of individuals or across contexts. These particular generalizations are pervasive in many empirical sciences and are one key part of scientific induction (Little, 1993; Slaney & Tafreshi, 2021). They will be the sole referent of the term 'scientific induction' here.

Scientific induction is usually held to higher epistemic standards than laypeople's everyday induction (Resnik & Elliot, 2016; Welch, 2018). Scientists, unlike laypeople, are typically expected and trained to assess the generalizability of their findings before extrapolating from them (Kukull & Ganguli, 2012; Engel & Schutt, 2013). Indeed, scientific induction is often judged by its "external validity", i.e., the extent to which inferences from a particular study context can be generalized to broader populations and across contexts (Shadish et al., 2002).

While much has been written on scientific induction (e.g., Cohen, 1970; Little, 1993; Sloman & Lagnado, 2005; Feeney & Heit, 2007; Achinstein, 2010; Claveau & Girard, 2019; Slaney & Tafreshi, 2021), one fundamental assumption has gone largely unquestioned. It is that scientific induction is an "act of reasoning" (Polit & Beck, 2010, p. 1451), a voluntary process that scientists fully control, rather than an automatic tendency that may often operate by default and affect scientists' research and conclusions against their epistemic goals (e.g., the formation of true beliefs).

Some accounts of human reasoning (such as dual-process views) do take induction to typically be a quick, unconscious heuristic process whereas deduction is construed as a deliberative, analytic process (Heit & Rotello, 2010; Evans & Stanovich, 2013; Stephens et al., 2018). Moreover, some studies found evidence that people's reasoning about individuals may involve an implicit ("system 1") generalization tendency (Sutherland et al., 2015). However, these notions have not yet been related to *scientific* induction and the "cognitive science of science", the interdisciplinary study of scientific thinking (Giere & Feigl, 1992; Thagard, 2012; Rich et al., 2021). Experimental work on inferential biases in general is "seldom considered when studying the behavior of research scientists" (Bishop, 2020, p. 3).[4]

[2] Induction is one of three major kinds of reasoning. The others are *deduction*, i.e., inferences from broad generalizations to conclusions about specific cases, and *abduction*, i.e., inferences to the best explanation (Walton, 2001). Here we focus specifically on induction because our argument is empirically best supported with respect to it.

[3] Many of our arguments also apply to science more generally. But since most of the evidence we will draw on comes from or pertains to the cognitive, behavioral, and social sciences, they will be our main focus.

[4] Exceptions include, for instance, Mahoney (1976), or Tversky and Kahneman (1971); we will return to their work below.

Moreover, drawing unreflective "default inferences" from particular samples to broader populations of individuals is commonly thought to be at odds with how scientists generalize (Claveau & Girard, 2019, p. 855). Relatedly, recent papers and textbooks on scientific methodology covering generalizability (e.g., Engel & Schutt, 2013; Yarkoni, 2020) do not consider that scientific induction may involve an automatic generalization tendency. That scientific induction is from the start an entirely voluntary process seems at least tacitly widely assumed among scientists.

Here we challenge this common view and in doing so aim to provide a new perspective in the cognitive science of science. By integrating research from across the cognitive sciences, we argue that scientific induction involves by default a *generalization bias*, i.e., a cognitive tendency that operates automatically and frequently leads researchers to unintentionally generalize their results from particular samples to broader populations including when their evidence does not warrant it and when the generalization is avoidable. The outcome, in scientific studies, is overgeneralizations, which are construed here (*inter alia*) as study conclusions that are formulated too broadly in light of the given data. We call this new view of scientific induction the *generalization bias account*.

The account does not imply that scientists' inductive generalizations are always beyond their control. Scientists make significant efforts to assess the generalizability of their results and commonly tailor their conclusions accordingly (Kukull & Ganguli, 2012; Engel & Schutt, 2013; Slaney & Tafreshi, 2021). The generalization bias account rather holds that scientific induction involves a default but modifiable extrapolation tendency that can and often does drive scientists to unintentionally overgeneralize. We will argue for this account by combining (1) insights on bounded rationality, (2) psychological data on generalization and explanation, (3) evidence of scientists' overgeneralizations from small and unrepresentative study samples, and (4) the benefits of the generalization bias account in explaining such overgeneralizations.

Our account has important implications. It suggests that many researchers' current thinking about the nature of scientific induction needs to be revised. The account also helps advance the debate on the "replication crisis" across the sciences, i.e., the challenge that many scientific results cannot be replicated (Nosek et al., 2021). To date, the literature on the causes of the replication crisis has predominantly focused on external factors tied to the methodological, social, economic, or institutional structure of science. This includes small samples, publication bias, *p*-hacking, etc. (Munafò et al., 2017). Some internal factors such as cognitive biases have been discussed as well (Bishop, 2020; Munafò et al., 2020). But while generalizability concerns have appeared in the replication crisis literature (Tiokhin et al., 2019; Yarkoni, 2020), the potential role of a generalization bias has not been considered. We argue that this bias too can contribute to replication failures. Explanations of this crisis that focus only on structural causes thus risk missing an important factor. Current difficulties in replicability and generalizability observed across science are best explained by combining structural and psychological accounts as we do here. Moreover, commonly proposed interventions to tackle scientific overgeneralizations and the replication crisis should be supplemented with debiasing strategies.

2. Terms and outline of the argument

When using the term 'bias', we do not mean to imply that the cognitive process referred to is necessarily problematic. Many cognitive biases can be adaptive (Gigerenzer, 2018). We use the term here only to indicate that the process can operate automatically (i.e., under cognitive

load; Shea & Firth, 2016), unintentionally, and in ways that lead to a systematic but in principle temporarily avoidable deviation from a canonical normative benchmark. Within the sciences, this benchmark may be an accurate description and interpretation of data (Shadish et al., 2002). This notion of 'cognitive bias' is consistent with the way the term is typically used in the cognitive sciences (Haselton et al., 2009; Stanovich, 2009; Kahneman, 2011).

Moreover, while we will mostly use the singular term 'generalization bias', there can be different generalization biases in human cognition. An automatic tendency to generalize (and so a generalization bias) about people might differ from an automatic tendency to generalize about objects with respect to its strength, context sensitivity, malleability, negative effects, or realizers within an agent's cognitive architecture. Thus, while generalization bias may reflect aspects of a domain-general tendency, the term here refers to a diverse phenomenon.[5] This pluralist bias notion aligns with the way in which terms such as 'confirmation bias' are used in the psychological literature (Hahn & Harris, 2014; Peters, 2022).

Our case for the generalization bias account of scientific induction involves three steps. We first introduce theoretical arguments and review empirical findings that suggest that human beings have a predisposition to generalize that can constrain the accuracy of their cognition, sometimes in avoidable ways, producing a generalization bias (section 3). Building on these points, we then argue that even scientists often exhibit such a generalization bias in scientific induction (section 4). After that, we outline the likely role of this bias in the replication crisis (section 5) and propose interventions to reduce the bias (section 6).

3. Bounded rationality and empirical evidence on generalization

Before focusing specifically on scientific induction, it is instructive to consider the tendency to generalize in human cognition more broadly. While there is extensive psychological research on human inductive reasoning (for an overview, see Hayes & Heit, 2018), the focus here will be on how inductive reasoning can lead to overgeneralizations, which is less explored. We first argue that a frequently adaptive predisposition to overgeneralize is driven by cognitive resource limitations, i.e., by the need to keep cognition tractable. We then link this predisposition to psychological data that support the existence of a generalization bias in everyday cognition.

3.1 Cognitive constraints and overgeneralizations

Suppose people routinely adjusted their inductive generalizations about the world to the evidence that they have so that their generalizations accurately capture reality. Take the claim *C*: 'Healthy nutrition increases people's life expectancy.' To capture all relevant factors pertaining to the truth of *C* would require the following rephrasing: 'Healthy nutrition increases our life expectancy *if* people do not have an intolerance with certain foods that are important to a healthy diet, *if* people do not face other biological constraints', and so on. To avoid any overgeneralization, we would need to keep adding qualifiers. But it is well known from research on bounded or computational rationality (Gershman et al., 2015) that this is not feasible in practice, as we only have limited cognitive resources (time, attention, memory capacity) available (Gigerenzer & Selten, 2001). As a result, the human cognitive system needs

[5] If there is not 'a' generalization bias but a host of generalization biases, to what extent do they share or differ in their functional roles across domains? Due to space constraints, we shall remain agnostic on this question. Our focus will be on the more fundamental question of whether this kind of bias is real in the domain of scientific induction.

to short-circuit its computations whenever possible, displaying a "strong bias to default to the simplest cognitive mechanism" (Stanovich, 2009, p. 64).

This matters for the notion of adding all relevant qualifiers to claim *C* to ensure accuracy. If humans are cognitive resource savers (Fiske & Taylor, 2013), then for human cognition to not get bogged down in computational intractability, the kind of envisaged restriction on the scope of *C* must at some point involve a trade-off between complexity (i.e., description length) and accuracy (Zaslavsky et al., 2018). To avoid computations whose adaptive costs will eventually outweigh the benefits, the human cognitive system will need to habitually opt for potential overgeneralizations. That is, we should expect the system to have a default disposition to curb its constraints in generalizing processes, producing overgeneralizations that are on average resource rational for the system in its environment of original adaptation (Lieder & Griffiths, 2019).

A more principled argument for this view can be derived from the "tractable cognition thesis", which states that any plausible theory of cognition must be computationally tractable such that information-theoretic considerations can be used to inform the study of feasible cognition (van Rooij, 2008). Given this thesis, how may, for instance, a scientist accomplish the cognitive task of accurately assessing the generalizability (i.e., the scope) of a theory or claim about people? Consider a case in which every individual in a population is assumed to differ by at least one qualifier (trait or condition). Individuals, in reality, have many distinct qualifiers including their genetics and environmental conditions. Any individual qualifier could thus be relevant to the truth of a given claim and thus would need to be assessed. Determining the precise scope would require assessing all possible allocations of these distinct individuals into two categories (within or without the scope of the theory/claim). For a population P we would thus have to assess 2^P cases. Given standard assumptions from the tractable cognition thesis (specifically, polynomial-time computability; Garey & Johnson, 1979), this means that the task quickly becomes computationally intractable ($P = 100$ individuals means considering $>10^{30}$ possibilities).

One way to make this problem tractable is to restrict our consideration to a small number k of relevant qualifiers (this has been called "fixed-parameter tractable" cognition; van Rooij 2008). The theory/claim may now apply to any subpopulation identified by a qualifier, or a combination of qualifiers from k. Assume we know the allocation of the k qualifiers in the population (age, sex, cultural specifics), but not which qualifiers are relevant. We then have to consider $2^k = f$ subpopulations instead of P individuals. To assess generalizability between sub-populations, we now have to consider only 2^f cases. While for a small k, the problem becomes tractable, for a growing k-value, a combinatorial explosion again quickly ensues, meaning that for a scientist it may only ever be possible to consider a smaller number of relevant qualifiers.[6] Formulating inductive claims with fully accurate scope is thus likely cognitively intractable in general. Conversely, a plausible tractable solution to this problem implies that the number of potential qualifiers that can be assessed is clearly limited, meaning that many potentially relevant qualifiers must be disregarded by default and it cannot be considered whether they are relevant. Yet, since any omission of a relevant qualifier constitutes an overgeneralization, the very need to keep cognition and scientific models tractable implies a fundamental and inevitable overgeneralization tendency.

[6] However, this does not preclude cumulative knowledge about qualifiers, which may, for instance, result from social criticism processes. Once a qualifier is known, it does not always need to be considered in each case and it does not necessarily reduce tractability.

We have focused on overgeneralizations resulting from cognitive limitations in *individual* cognizers. However, cognitive limitations of individuals can become compensated in groups or be corrected in social exchanges with others (Dutilh-Novae, 2020, pp. 151-167). Later we will explore to what extent this also applies to overgeneralizations in individuals' cognition (section 4.5). But first we turn to generalization bias itself.

3.2 Empirical data on generalization bias

The fundamental disposition to overgeneralize that we just outlined is inevitable for computational reasons. But it may also lead to a separate automatic tendency to generalize, one that operates by default but can (with some attention, interest, etc.) be suspended. That is, it may lead to a generalization bias. This bias can be illustrated by reference to people's thinking in terms of kinds, i.e., whole categories of beings or objects ('women', 'trees', etc.), rather than only in terms of unique individuals or objects (e.g., '*this* woman', *'this* tree'), or explicitly quantified sets of them ('*some* women', '5 trees') (Pelletier, 2009). Research suggests that human cognition may privilege and prefer information processing at the level of kinds.

Consider studies on people's understanding of *generics*, i.e., explicitly unquantified claims expressing generalizations about kinds of individual beings or objects (e.g., 'mosquitos carry malaria', 'smartphones are popular'; for discussion, see Leslie and Lerner, 2016). Researchers found that when children and adults were asked to recall both generics about animal kinds and quantified statements about them, they were more likely to recall quantified statements as generics than do the reverse (Leslie & Gelman, 2012).

In a related experiment, Sutherland et al. (2015) asked participants to correctly memorize statements about novel fictional[7] animals. The statements were either quantified 'many' claims about individual agents, or generics about groups of them. On average, 63.5% of the 'many' claims were misremembered as generics, and there were significantly more (M = 57.4%) 'many'-to-generic than generic-to-'many' conversions. The rates at which participants misremembered quantified claims also did not significantly differ in a cognitive-load condition vs. a no-load condition, suggesting that participants had a "generalization bias",[8] an "implicit bias to spontaneously generalize to kinds" (Sutherland et al., 2015, p. 1038). The participants' responses indicate a *bias* because these overgeneralizations happened systematically and despite individuals' attempts and ability to correctly encode information (it doesn't follow that if many *Xs* are *F*, then *Xs* in general are *F*): Participants could in principle avoid incorrect answers (i.e., doing so was not cognitively intractable) but nonetheless unintentionally and reliably deviated from this normative standard.

Indeed, people may need only very little evidence about some individuals to readily generalize to entire kinds of them: Cimpian et al. (2010) introduced their participants to fictional animals ('lorches'), before telling them that a certain percentage of category members have a particular feature (e.g., '30% of the lorches have feature *F.*'). Then they asked participants if the corresponding generic (e.g., 'Lorches have feature *F.*') was true. Participants treated the generic as true even when they knew that only a minority of the kind members had the feature. Yet, when participants were *first* told that this generic statement (i.e., 'Lorches have feature *F.*') was true and were then asked what percentage have feature *F*, participants expected over

[7] Fictional animals were used to prevent individuals' prior knowledge about the world from influencing their responses.

[8] To the best of our knowledge, Sutherland et al. (2015) were the first to use the term "generalization bias" but they employed it primarily only to refer to a memory bias.

90% of the individuals at issue to have *F*. These data reveal an unreflective tendency to generalize vastly beyond the evidential basis. Different findings thus suggest that the human cognitive system has an automatic generalization bias that operates by default and often results in unwitting, avoidable overgeneralizations.

That said, Cimpian et al.'s, and Sutherland et al.'s studies used only Western student samples and statements about fictional kinds. Also, consider the fact that social stereotypes, which themselves indicate a tendency to overgeneralize from some to all members of a kind, cannot easily be overcome by presenting people holding them with numerous counterexamples (Hinton, 2017). If people had an automatic tendency to generalize that is context-independent, people should also readily generalize from some counter-stereotypical individuals to the corresponding kind of them. Since this does not commonly happen, the impact of generalization bias is likely influenced by psychological, contextual, and social factors (e.g., motivated reasoning; Kunda, 1990). These qualifications aside, the reviewed findings and the preceding bounded rationality argument do provide reasons to believe that many people have a generalization bias, a fundamental, automatically operating extrapolation tendency that leads them to unintentionally and avoidably generalize their views about particular cases to broader sets (e.g., kinds of individuals) even when their evidence does not support it. Does this bias also affect *scientific* induction?

4. Scientific induction

Scientific methodology has partly been developed to combat unwarranted generalizations and ensure generalizability. For instance, inferential statistics is designed to systematically assess the generalizability of data from a sample to a larger population (Kukull & Ganguli, 2012). Moreover, science involves researchers routinely criticizing each other's inferences (e.g., in peer review) to ensure reliable scientific belief formation (Longino, 2002). Additionally, comparative research on induction found that domain experts often make inductive inferences based on their deeper domain knowledge rather than on general heuristics (Hayes & Heit, 2018, p. 3). It might thus seem that scientists are unlikely to display generalization biases in scientific induction. However, we now argue otherwise and make the case for the generalization bias account of scientific induction. To do so, we will integrate a range of different empirical findings and theoretical insights that have so far not been connected. Taken together, and combined with the arguments from the previous section, they yield a convincing overall case for the generalization bias account.

4.1 Scientific explanation can promote an overgeneralization tendency

There is reason to believe that scientific induction can in fact be particularly likely to involve a generalization bias. Specifically, this bias may be promoted by the motivation and practice of providing explanations, which is pervasive in science. Consider first work on laypeople's explanations of why events unfold in particular ways, why people behave as they do, etc. Studies show that when learners are asked to explain events, they "learn more effectively and generalize more readily to novel situations": "explaining guides learners to interpret what they are learning in terms of unifying patterns or regularities", which facilitates "broad generalizations" (Williams & Lombrozo, 2010, p. 776). Clearly, the mindset activated by being prompted to explain a phenomenon has thus benefits (it boosts learning).

However, it can also produce systematic *errors*. Williams et al. (2013) asked participants to categorize new objects (e.g., cars) after an exemplar-based training phase. During the training,

some individuals had to explain the categorizations. The others were asked to simply 'think aloud' about the task. Williams et al. found that the explanation group more accurately categorized features and objects that had similar patterns to the training examples, but less accurately categorized exceptional cases and ones with unique features. That is, "explainers focused on features that supported patterns at the expense of idiosyncratic information about individual items, and [...] perseverated in seeking or applying broad patterns despite evidence against their generality", producing "overgeneralization in the face of exceptions" (Williams et al., 2013, p. 1006). Relatedly, research on stereotyping found that asking participants to explain a single (unrepresentative) observation produced the same type of overgeneralizing beliefs that underlie social stereotypes, which ascribe properties to whole classes of individuals based on few observations (Risen et al., 2007).

These studies did not sample scientists. But it is a key part of science to explain, predict, and confirm hypotheses about phenomena (Kitcher, 1989). Since scientists (unlike laypeople) are as part of their job routinely aiming to explain phenomena (Cummins, 2000), they may be particularly prone to the kind of overgeneralizations in the face of exceptions that Williams et al. (2013) discovered. After all, many scientists also view "explanatory unification" (Kitcher, 1989), the project of explaining much by little and reducing the number of apparently independent phenomena, as a "virtue to be pursued in scientific theorizing" (Mäki, 2001, p. 488). This increases the plausibility of the view that scientists are especially likely to also display the kind of tendency to overgeneralize found in laypeople. Chomsky (1959) mentions a classic example of such an overgeneralization in the face of exceptions in (early) cognitive science: behaviourists took the idea that *some* behavior is conditioned to apply to all behavior.

Additionally, research suggests that in explaining events, people often generalize information in proportion to how salient or noteworthy it is (for a review, see Leslie, 2017). Cimpian et al. (2010) found that even if people learned that only few members of a kind had a certain feature (e.g., 10%), if that feature was distinctive or unusual, people more readily unreflectively overgeneralized the possession of the feature to the entire kind of individuals than if it was not distinctive. This study too did not sample scientists. However, scientists routinely try to extend the domain of what is already known. They should therefore often encounter still unknown, thus unusual, features of (e.g.) people and indeed actively seek striking, impactful findings (West & Bergstrom, 2021). Taken together, these points suggest that scientists may be at a particularly high risk of being affect by a generalization bias.

4.2 The belief in the law of small numbers

There is more direct empirical evidence of generalization bias in scientific induction. In a seminal study, Tversky and Kahneman (1971) found that many psychologists viewed a sample randomly drawn from a population as highly representative, i.e., generalizable, even when this was not warranted. Most of the surveyed psychologists underestimated the systematic increase in uncertainty for smaller samples such that they placed about the same confidence in a mean derived from a small sample as in a mean derived from a larger, more representative sample.

In statistics, the "law of large numbers" is a foundational theorem stating that the mean of a sample provides an increasingly reliable estimate of the mean of the population as the sample size increases (Dekking, 2005). Inversely, estimates from smaller samples tend to be less reliable. Tversky and Kahneman (1971) thus dubbed the systematic overconfidence in estimates obtained from small samples the "belief in the law of small numbers". With respect to this "belief" among psychologists, it is fair to assume that the psychologists that Tversky and

Kahneman surveyed did not *deliberately* overgeneralize. Moreover, they should not have (given their professional expertise)[9] lacked the competence required for accurate responding. More plausibly, Tversky and Kahneman's findings indicate that their participants were affected by an automatically operating generalization tendency.

We have no reason to believe that scientists today are immune to the belief in the law of small numbers. In fact, recent studies found that, across the cognitive and behavioral sciences, published sample sizes are persistently too small to support the purported conclusions (Button et al., 2013; Smaldino & McElreath, 2016; Szucs & Ioannidis, 2017). There are likely multiple convergent causes of this phenomenon including institutional and structural causes that we consider below (section 4.5). But Bishop (2020) argues convincingly that the belief in the law of small numbers plays a role as well. Since this belief indicates an unreflective tendency to overgeneralize from the mean of a given study sample to the mean of the total population, it indicates a generalization bias. The persistence of insufficient sample sizes across the cognitive and behavioral sciences then suggests that this bias is present in many cases of scientific reasoning.

4.3 Extrapolating from unrepresentative samples

We just illustrated one important way in which generalization bias can affect scientific induction, namely by leading scientists to assume that their current finding for a small sample is representative of the population from which the sample is drawn. This can influence statistical inferences resulting in overgeneralizations. We argue next that the bias is likely to also affect scientific induction in another kind of inference, namely when scientists extrapolate to *other* populations[10] than those from which their sample is drawn. This may result in overly broad claims about humans in general.

Such overgeneralizations can in fact frequently be encountered in academic publications. For instance, in an influential study, Henrich et al. (2010) found that:

> Behavioral scientists routinely publish broad claims about human psychology and behavior in the world's top journals based on samples drawn entirely from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) populations. Researchers – often implicitly – assume that either there is little variation across human populations, or that these 'standard subjects' are as representative of the species as any other population. (p. 61)

Accordingly, researchers "often implicitly" overgeneralize. This is because upon reviewing a comparative database from across the behavioral sciences, Henrich et al. found that "WEIRD subjects are particularly unusual compared with the rest of the species – frequent outliers" (2010, pp. 64-78). An explicit consideration of generalizability should thus have prevented broad claims about human psychology based on purely WEIRD samples.

While Henrich et al. did not quantify these kinds of overgeneralizations, Rad et al. (2018) did. They reviewed 223 articles published in *Psychological Science* (from 2014-2017) to evaluate the extent to which studies in the field rely on sampling WEIRD populations and make inferences

[9] The sample was drawn from meetings of the Mathematical Psychology Group and of the American Psychological Association – with generally all psychologists having training in statistics.

[10] The term 'other' here can be defined as a difference in at least one potentially relevant qualifier (e.g. cultural background); see section 3.1.

from them about humans in general. They found that ~94% of the studies (total $N = 447$) sampled only people in Western countries, and 76% contained generalized conclusions. That is, only 24% contained conclusions limited to the WEIRD sample (see Rad et al.'s Table S2, Supplementary Information).

Similarly, DeJesus et al. (2019) analyzed 1,149 psychology articles (published 2015-2016 in 11 journals) to investigate the use of generics in titles, research highlights, and abstracts. As noted, generics make broad claims about a category as a whole (e.g., 'people', 'introverts', 'women', etc.) vs. particular individuals (e.g., 'the people in the study') and don't mention frequencies, probabilities, or statistical distributions. DeJesus et al. found that "generics were ubiquitously used to convey results (89% of articles included at least 1 generic)" (2019, pp. 18370). Yet, most articles didn't mention sample demographics, and there was "no evidence that [the use of generics] was warranted by stronger evidence, as it was uncorrelated with sample size": "authors showed an overwhelming tendency to treat limited samples as supporting general conclusions, by means of universalizing statements" (ibid). In their papers, researchers thus very frequently overgeneralized.

Evidence of overgeneralizations does not only come from expressions of scientists' inferences from human (small or WEIRD) samples. Leavens et al. (2010) found, that the "over-reliance in psychology on one group of humans, WEIRD, to represent 'the human' in cognitive terms has a strong parallel in the over-reliance in comparative psychology on chimpanzees raised in *B*arren, *I*nstitutional, *Z*oo, *A*nd other *R*are *R*earing *E*nvironments (BIZARRE) to represent 'the chimpanzee'" (p. 100).

It might be objected that when scientists generalize from small, WEIRD, or BIZARRE samples to kinds and use generics in communicating scientific results, their generalizations are not in fact *over*-generalizations. This is because generic scientific generalizations are *ceteris paribus*, carrying an implicit 'all other things equal' caveat. That is, they are not meant to hold universally but allow for many exceptions (Claveau & Girard, 2019).

However, if scientists did indeed typically understand generic scientific generalizations as only referring to a subset of people, for instance, WEIRD populations, there would be little ground for Henrich et al. (2010), DeJesus et al. (2019) and many others to criticize researchers for these generalizations. They would simply not be taken to be false or problematic to begin with. Yet, they are commonly viewed as problematic (e.g., by the vast majority of the respondents to Henrich et al.'s paper; see Henrich et al., 2010, p. 51). These points suggest that many scientists do not usually implicitly relativize the generic generalizations at issue to WEIRD populations, and so these generic claims are in fact overgeneralizations.[11] Relatedly, Haigh et al. (2020) found that people high in "cognitive reflection" (the tendency to reflect on and revise

[11] Scientists who generally agree with the *ceteris paribus* view may nevertheless take the overgeneralization from WEIRD sampling to be impermissible. Closer scrutiny as to why this might be reveals that it hinges on the question as to what exactly is meant by 'ceteris'. The *ceteris paribus* language derives from model-based science in which 'all else' can be held constant. In empirical contexts, however, there are always many differences between two studies (context, stimuli, etc.). To make sense in empirical contexts, *ceteris paribus* needs to refer only to the assumed substantively relevant qualifiers (see section 3.1). For instance, the date, exact identities of participants, etc. must explicitly *not* fall under 'ceteris'. But identifying the relevant qualifiers is precisely the problem. Claiming that an implicit *ceteris paribus* justifies the omission of mentioning, for instance, that the sample is WEIRD is thus a universal cop-out: It can be invoked whenever any claim is not reproducible or generalizable for any reason. The crucial question is if readers understand the potential relevance of the sample being WEIRD in cases where sample identity is not explicitly mentioned. To the extent that they do not, the omission induces an overgeneralized interpretation.

one's intuitions), which perhaps includes trained scientists, did not display more restricted interpretations of science-related generics than other people.

What might explain the apparent pervasiveness and persistence (Nielsen et al., 2017; DeJesus at al., 2019) of overgeneralizations in the sciences? It seems clear that many of the researchers that produce them do not wish to deliberately misrepresent their findings. It might be that researchers sometimes overgeneralize because they do not yet fully understand the parameter space of a discovery, or have mistaken preconceptions about the phenomenon studied. This can result in inevitable overgeneralizations typical of normal science (Guttinger & Love, 2019).

However, mere lack of knowledge seems an insufficient explanation in the present cases. The point that WEIRD samples tend to be outliers has been emphasized and widely acknowledged for some time (Henrich et al., 2010). And assessing the degree to which data can be generalized is a basic skill taught in statistics courses and social science textbooks (Shadish et al., 2002; Engel & Schutt, 2013). The findings of consistent scientific overgeneralizations (e.g., from WEIRD or small samples) are thus more plausibly explained by holding that many scientists are affected by an unintentionally operating tendency to generalize. Some researchers' self-reports further support this. For instance, DeJesus et al. (2019) note that when they were conducting their research: "we were chastened to discover *unintended* generics in our own published writing [emphasis added]" (p. 18375). Similarly, consider organizations such as the American Psychological Association, which often release policy statements regarding science and behavior and are explicitly committed to providing accurate scientific information. When Elson et al. (2019) examined these organizations' policy statements (produced by scientists) they found problematic overgeneralizations (i.e., claims which extended research findings to behaviors beyond what was appropriate, e.g., when lab measures of aggression were readily extrapolated to real-world violence) in 62.5% of them. Since these overgeneralizations are highly unlikely to be intentional, there are reasons to believe that they are partly based on, and indicative of, a generalization bias.

4.4 Overgeneralizations in study designs

Having highlighted overgeneralizations from small or WEIRD participant samples, generalizability issues can in fact also arise for any other dimension of a study over which the researchers wish to generalize, for example, stimuli, situations, experimenters, research sites, etc. (Brunswik, 1947). Focusing on stimuli, across experimental psychology studies, participants are regularly asked to respond to items assumed to capture categories of theoretical interest. For instance, in implicit cognition tasks, participants might be primed with pictures of specific elderly vs. adolescent faces sampled in some way from other available stimuli. There could be significant stimulus variability, as items to prime people might be photos, videos, face-to-face exposure, etc. For their conclusions to generalize to other stimuli and not just to the specific ones used in their study, researchers would thus need to routinely account for the representativeness and variability in stimuli in both theorizing and statistical modelling. In reality, however, many studies found that stimulus sampling is rarely factored in during experimental design and data analysis (Clark, 1973; Judd et al., 2012; Yarkoni, 2020).

If scientists were immune to generalization bias in their scientific induction, one would expect that they are aware of the related generalizability problems and either (a) ensure that their stimuli are representative, or (b) explicitly report limited generalizability. However, very often neither happens. For instance, while psycholinguists rarely offer support for the view that their findings generalize beyond the specific sample of language materials they selected, they

frequently conclude that their findings hold for language in general (Clark, 1973; Judd et al., 2012). Yarkoni (2020) shows that many psychologists draw similar, overgeneralized inferences by not controlling stimuli as random factors and thus tacitly assuming that stimulus effects remain invariant across settings.

Different explanations for researchers' failure to attend to stimulus variability have been proposed. For instance, traditional analysis of variance procedures create high demands when both participants and stimuli are treated as random factors (Judd et al., 2012). Others view this failure as an accident of history, or of technological limitations since the computing resources needed to fit the required models were difficult to attain until recently (Yarkoni, 2020).

However, these proposals leave it unclear why researchers still persistently omit stimulus-related qualifiers in their reporting. Lack of knowledge of the related generalizability problems is an insufficient explanation: It is typically part of psychologists' first course in experimental design to learn that stimuli (situations, etc.) can influence participants' responses and that precautions thus need to be taken (Sani & Todman, 2006). That researchers intentionally misrepresent their results is also unlikely. It is more plausible that the overgeneralizations in study designs (including the 'stimulus-as-fixed-effect fallacy', Clark, 1973) and subsequent reporting are based on an automatic generalization tendency that inclines scientists to unwittingly disregard variability and qualifications in their research design and conclusions (e.g., to keep scientific cognition tractable). That is, despite thorough scientific training, researchers are likely to have a generalization bias.

By integrating psychological work on explanation (e.g., Williams et al., 2013) with a range of disparate findings from research about scientists themselves (Tversky & Kahneman, 1971; Clark, 1973; Henrich et al., 2010; DeJesus et al., 2019; Yarkoni, 2020), we thus arrive at a new perspective in the cognitive science of science: the view that scientific induction involves a generalization bias. This view helps unify the findings and cast new light on the nature of scientific induction.

4.5 Psychological and structural factors interact

There is a natural response to the preceding argument: External factors belonging to the economic, social, or institutional structure of science can also help explain the phenomena we just appealed to in order to support the generalization bias account. For instance, the economic costs of studies with larger samples, publication bias, methodological training, or academic competition also contribute to the use of small, WEIRD samples, and overgeneralizations from them (Smaldino & McElreath, 2016; Higginson & Munafò, 2016; DeJesus et al., 2019; Braganza, 2022).

We will now argue that such structural causes of the problems related to sampling and generalizability should not be misconstrued as explanatory factors *competing with* generalization bias in accounting for these problems. Instead the psychological and structural drivers of small, WEIRD study samples and scientific overgeneralizations (e.g., in study designs) mutually reinforce each other. Both explanatory approaches are thus best combined.

Consider a structural explanation of why scientists might frequently select small and unrepresentative samples and overgeneralize from them in publications, namely that this practice is in part driven by publication pressure (Braganza, 2020). We have three principal ways to explain the scientists' mindset in these cases. We might hold that (1) the researchers

are completely unaware of the issues related to overgeneralizations from such samples. (2) They may be familiar with some of them but nonetheless consciously select too small samples to advance their career. Or (3) they may remain largely unaware of overstating their findings, as they are affected by a generalization bias.

Since the problems concerning small samples and overgeneralizations from them have been highlighted in the cognitive and behavioral sciences for decades and are widely acknowledged (e.g., Henrich et al., 2010; Smaldino & McElreath, 2016), (1) seems an insufficient explanation. As for (2), while scientific misconduct exists (John et al., 2012), most scientists are unlikely to be consciously complicit in what they perceive as scientific malpractice. In our view, option (3) is most convincing. The assumption that scientists are affected by a generalization bias makes it easier to see how structural factors (e.g., publication pressure) can cause and promote the selection of small, unrepresentative samples and overgeneralizations from them. Indeed, since scientists are unlikely to be either oblivious to these problems or uninterested in their epistemic consequences, the assumption that a bias is at play makes structural accounts (e.g., Smaldino and McElreath's (2016)) significantly more plausible. The generalization bias account thus helps explain how scientists may be affected by structural causes of the problems at hand.

The reverse holds too, because structural factors, in turn, can promote generalization bias. It is rarely feasible in science to collect data on an entire target population, or have a fully representative sample (Banerjee & Chaudhury, 2010). While many scientists aim to select a sample that is representative of their target population, they often face external constraints beyond their control. For instance, stratified random sampling can require significant resources, people contacted for a study may not want to participate, or the relevant target population may not be fully known (Martinez-Mesa et al., 2016; Tyrer & Heyman, 2016). Even though having a fully representative sample is thus rare in science, in communicating their findings to policy makers, scientists nonetheless often need to make broad generalizations when they provide explanations and recommendations (Lombrozo, 2013; Peters, 2021): Policy makers and the public want to know whether *they*, not just the participants in a particular study, should eat food rich in antioxidants, etc. Social and structural factors therefore promote overgeneralizations by encouraging researchers to use bolder framing when they need to persuade journal editors, funding agencies, and the public of the importance of their research (Guttinger & Love, 2019; DeJesus et al., 2019). Relatedly, given standard statistical practices, small sample sizes can allow for more publishable, i.e., positive results, providing scientists with an advantage in competition for funding and academic positions while simultaneously undermining external validity (Braganza, 2020). This would structurally reinforce the "belief in the law of small numbers" (Tversky & Kahneman, 1971) by selecting the scientists displaying it for tenure (Smaldino & McElreath, 2016).

Social and structural factors can thus incline scientists to make, and think in terms of, overgeneralizing claims, instilling and strengthening habits (Verplanken, 2018) that may subsequently operate by default and result in generalization bias. Consequently, both generalization bias and these structural factors will need to be considered together to account for scientific overgeneralizations. This approach fits nicely into, and adds a novel aspect to, the increasing number of contributions across disciplines that highlight the importance of connecting psychological and socio-structural factors in explaining and addressing complex social challenges (e.g., the replication crisis: Munafò et al., 2020; social injustice: Davidson & Kelly, 2020; climate change: Brownstein et al., 2022).

But given our emphasis on the interplay between psychological and social factors, an important question remains. We noted above that science involves researchers routinely criticizing each other to reduce the influence of an individual scientist's biases (Longino, 2002). Might peer review, social criticism, etc. not also help keep generalization bias in the sciences in check? Perhaps they do to some extent. However, the fact that overgeneralizations from, for example, small or WEIRD samples are *pervasive* in many top science journals (DeJesus et al., 2019) strongly suggests that the corrective power of these social feedback mechanisms is rather limited: Scientists do not only routinely produce overgeneralizations but also routinely *let them through peer review*. That is, in many cases, even peer reviewing scientists seem to have an unreflective overgeneralization tendency when considering their colleagues' work. This indicates that the current social criticism mechanisms are frequently ineffective against overgeneralizations and, by extension, generalization bias in science. Indeed, since peer reviewers have the social function and usually explicit goal to rigorously examine manuscripts to detect errors (including unwarranted claims), the common oversight of problematic overgeneralizations during peer review is likely unintentional, suggesting that generalization bias can have powerful effects even among highly epistemically vigilant scientists.

This completes our case for the generalization bias account of scientific induction. To avoid falling prey to a generalization bias ourselves (see also work on "bias bias"; Gigerenzer, 2018), we provided a wide range of different kinds of evidence and insights that support the account. They include arguments from bounded rationality and tractable cognition, psychological data on generalizations (section 3), findings on explanation and scientific overgeneralizations (sections 4-4.4), and the explanatory benefits that the account yields.

We did not argue that scientists are always negatively affected by a generalization bias in their scientific induction. We grant that they often do not automatically generalize in unwarranted ways but carefully scale their claims to the evidence. Our point is that scientific induction involves an implicit, unreflective generalization tendency that is operative by default but can be and often is mitigated when researchers have cognitive resources available. This is in line with the dual-process view of reasoning (Evans & Stanovich, 2013), which has not been systematically brought to bear on scientific cognition yet,[12] and the thought that the bias at issue is influenced by psychological, social, and contextual factors (e.g., motivated cognition, Kunda, 1990).

Finally, our account allows that generalization bias may sometimes result in *appropriate* generalizations. While we shall not delve into a normative analysis of inductive support in science (e.g., Cohen, 1970; 1988), whether an instance of scientific induction is negatively affected by an automatic generalization tendency depends partly on the extent to which researchers' study samples (including participants, stimuli, situations, etc.) are representative of the larger populations they are drawn from. As illustrated in Figure 1, the smaller and less representative the sample, the higher the risk of generalization bias producing generalizations that lack sufficient evidence, i.e., overgeneralizations.[13] Since scientists' study samples are

[12] But for relevant developmental psychological work, see Amsel et al., (2008).

[13] Whether a generalization is an *over*-generalization depends on the background knowledge (context, etc.) of the receiver; domain expertise may lead some individuals to add relevant scope-restrictors to a given generalizing statement that novices would not add.

always on a scale from smaller and less representative to larger and more representative,[14] the likelihood of generalization bias resulting in overgeneralizations is always on a spectrum as well.

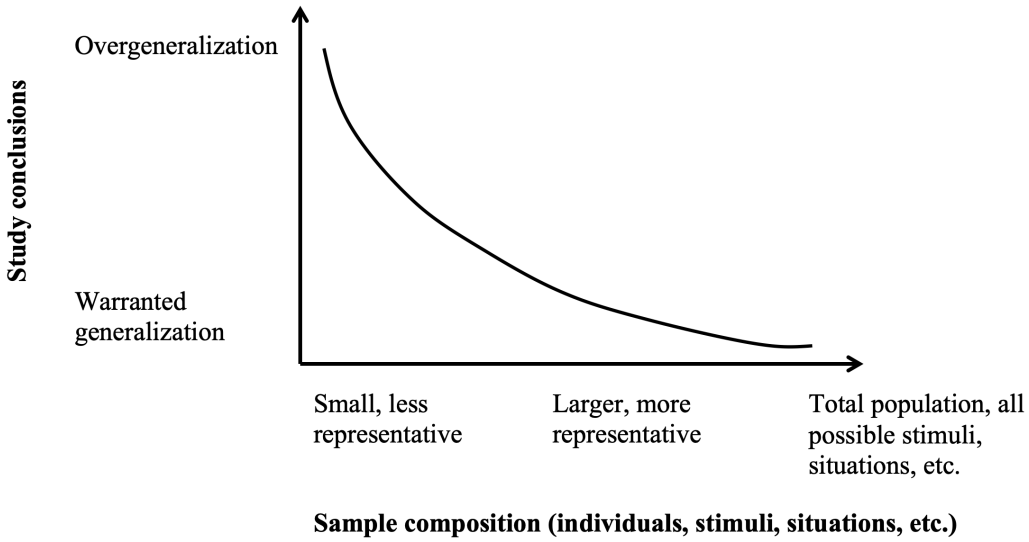

Figure 1. The relationship between sample composition (size and representativeness) and the likelihood of overgeneralized study conclusions

5. Generalization bias can contribute to the replication crisis

Many sciences are thought to be facing a replication crisis (Baker, 2016). A systematic replication project in psychology found that while 97% of the original studies assessed had statistically significant effects, only 36% of the replications yielded significant findings (Open Science Collaboration, 2015). Much has been written on possible causes of replication failures (Lewandowsky & Oberauer, 2020). Munafò et al. (2017) argue that replicable science is threatened by poor study design, hypothesizing after the results are known, low statistical power, *p*-hacking (i.e., manipulating statistical analyses until statistically significant results emerge), lack of data sharing, publication bias, and many other factors.

Several researchers have argued that the replication crisis is closely connected to a "generalizability crisis" in the sciences (Redish et al., 2018; Tiokhin et al., 2019; Yarkoni, 2020). However, generalization bias has not been considered in this context yet. The account of scientific induction we developed above provides a new insight into some underlying causes of the replication crisis. For example, Yarkoni (2020) argues that overgeneralizations in study designs due to oversight of stimuli or situation variability contribute to replication failures: Since a minor change in the uncontrolled stimuli or situational factors that researchers generalize over can often "take researchers from $p = .5$ to $p = .0005$ or *vice versa*", replicating the particular effects of previous studies becomes problematic and so researchers' overgeneralizations facilitate replication difficulties (Yarkoni, 2020, p. 17). While Yarkoni does not consider psychological contributors, generalization bias can significantly exacerbate the problem by feeding into these overgeneralizations.

[14] The question of whether generalizations derive more support from the variety of instances that are cited as evidence for them or from the multiplicity of those instances is a matter of philosophical debate (see, e.g., Cohen, 1988).

Furthermore, it has been argued that some overgeneralizations that are inevitable in science can contribute to replication failures too (Guttinger & Love, 2019). When researchers (1) do not consider or fully understand all variables affecting the phenomenon they study, or when they (2) have mistaken preconceptions about the phenomenon they study then partial models that overlook relevant factors and connections may result. Overgeneralizations based on (1) or (2) are to be expected because when scientists approach the limits of knowledge (as they often do), they inevitably do not fully understand all of the parameter space for, and qualifiers of, a phenomenon, and so cannot control for them in their study designs (see section 3.1). This can affect not only aspects of external validity but also some aspects of internal validity (e.g., the control of potential confounders in study design to ensure certainty that the independent variable influenced the dependent variable). The argument we developed above offers additional insight on this point. It introduces a causal factor that may impair scientists' exploration of the parameter space of a discovery or support mistaken preconceptions that may then result in overgeneralizations: Generalization bias can incline researchers to avoid making a lack of understanding explicit, or to not revise their misconceptions. This exacerbates the impact of (1) and (2) on scientific induction.

6. Interventions

A generalization bias in scientific induction has important consequences for the question of how to reduce problematic generalizations in study designs and conclusions. Researchers aware of the problems associated with small and WEIRD samples and overgeneralizations from them have developed different measures for tackling them (Munafò et al., 2017). The generalization bias account helps inform and extend these measures.

One central recommendation to reduce scientific overgeneralizations is to improve journals' author guidelines: Journals should include, among their existing reporting requirements, a requirement for authors to provide "constraints on generality" statements in each paper, which should be outlined in the methods and conclusion sections (Simons et al., 2017, p. 1123). These statements should justify the authors' claims about a study's design and target populations, the particular characteristics of participants, stimuli, procedures, and other contextual features that allow others to assess the results' generalizability. While authors do already sometimes discuss the generalizability of their results in their papers' 'limitations' sections, the use of specific reporting requirements to ensure that this happens routinely is still not established practice across scientific journals (Moher et al., 2012; Yarkoni, 2020).

The generalization bias account adds weight to calls for adopting reporting requirement to be explicit about generalizability. This is not only because constraints on generality statements benefit prospective readers, but also because they encourage authors to routinely and more explicitly reflect on the appropriate degree of generalizability of their findings. Such conscious deliberation is a first line of defence against cognitive biases, including generalization bias.

However, structural interventions such as improved reporting guidelines explicitly concern only publication practices while overgeneralizations can also occur in other contexts (e.g., teaching) where reviewer oversight is absent or peer criticism limited. Moreover, if scientific overgeneralizations often arise due to a generalization bias then what is needed in addition to improved author guidelines is cognitive "debiasing", i.e., interventions that directly target scientists' cognition (Larrick, 2004; Croskerry et al., 2013). This is especially important given that even peer reviewers are frequently affected by a strong generalization tendency, as evidenced by the high number of overgeneralizations in published (i.e., peer-reviewed)

scientific papers (DeJesus et al., 2019). Cognitive debiasing is already being used in the clinical and forensic sciences to minimize scientists' cognitive biases (e.g., confirmation bias, Lockhart & Satay-Murti, 2017) and improve diagnostic accuracy (Daniel et al., 2017; Ludolph & Schulz, 2018; Sibbald et al., 2019). But the idea of using it in the cognitive and behavioral sciences is largely unexplored (Bishop, 2020). We think that it should be viewed as equally important in these sciences, and relevant interventions should be developed (Morewedge et al., 2015; Sellier et al., 2019).

To take a step in this direction, two straightforward and concrete debiasing measures come to mind that can also make structural interventions against overgeneralizations more effective:

(1) *Raising awareness*. Cognitive debiasing begins with individuals' awareness of their own biases. This can already help mitigate them (Croskerry et al., 2013). To illustrate this in the context of generalization bias, suppose that to minimize the risk of overgeneralizations, a group of scientists follows the journal reporting guidelines outlined above. If they did not only follow these guidelines but also thought that they might unintentionally and automatically overgeneralize from their samples, this should increase their attention to scaling their claims to the data, which in turn minimizes the influence of generalization bias, reduces overgeneralizations, and boosts the impact of reporting guidelines. Introductory methodology courses at universities and author guidelines should thus raise people's awareness of generalization bias.

(2) *Checklists.* The efficacy of raising awareness may be limited, as biases are known to be recalcitrant to conscious control (Kurdi & Banaji, 2021). That is why researchers working on debiasing often recommend the use of personal checklists (Lockhart & Satay-Murti, 2017), an analogue to journal checklists and guidelines. A relevant checklist that individual scientists can use to control and over time minimize their own generalization biases includes three sets of items. One would cover common sources of variability specific to the chosen study type (participants, stimuli, etc., see the CONSORT guidelines; Moher et al., 2010). Another set would capture the stages of the research process at which an overgeneralization overlooking these factors can occur under the influence of an automatic generalization tendency (e.g., study design). A third set would concern science reporting (e.g., in articles), require clearly outlining specific relevant qualifiers (e.g., on sample size and composition), and recommend using the past tense when reporting results (DeJesus et al, 2019). This communicates that the findings pertain to the sample or analysis at hand and may thus help undercut potential overgeneralizations by the reader. If scientists adopt such checklists consistently and not only during the publication process when reviewers prompt them, they will more widely expose themselves to an overgeneralization "habit-breaking" regime (Forscher et al., 2017). This can promote adherence to the structural interventions outlined above. Structural and psychological interventions should thus be combined to more effectively minimize scientific overgeneralizations.

The robustness and replicability of science can also be increased by conducting a single study in multiple contexts, and applying multiple methods and evidence from different fields to better assess the scope of a given finding (Nosek et al., 2021). Other structural mitigation strategies from the rapidly expanding literature on irreproducibility that target different scientific stakeholders include pre-registration, results-occluded peer review, improved randomization, blinding procedures, and data availability requirements (Munafò et al., 2017). However, such measures do not directly address generalization bias but rather reproducibility and

generalizability more broadly. We will thus not further discuss them here but note only that a number of them are likely to interact with generalization bias in complex ways.

Finally, it is worth emphasizing that the perspective that we are defending here is valuable not just in suggesting new ways to effectively intervene on the problems of overgeneralizations, WEIRD sampling, replication failures, etc. It also changes our understanding of these problems themselves. Properly understood, we need to *reconceive* the nature and source of overgeneralizations, WEIRD sampling, and replication failures. The generalization bias account of scientific induction gives a richer diagnosis of the underlying causes, and how they are located in a set of structural features, psychological features, and interactions between them.[15]

7. Conclusion

Scientific induction is commonly treated as a fully voluntary act of reasoning. Accordingly, unreflective default inferences from a given sample to a broader population are taken to be at odds with how scientists generalize. Here we challenged this view and developed an alternative. We argued that scientific induction involves by default a generalization bias that operates automatically and frequently leads researchers to overgeneralize their results unintentionally. Research on bounded rationality and psychological evidence on generalization suggest that generalization bias is a property of everyday induction. Moreover, findings on explanation and scientific overgeneralizations suggest that this bias also affects scientific induction. The explanatory benefits of the generalization bias account in making structural explanations of overgeneralizations more plausible lend further support to this account. Combined with the points made earlier in the paper (on bounded rationality and the evidence on generalization), these arguments yield a compelling overall case for the generalization bias account.

This account provides grounds to hold that generalization bias contributes to the replication crisis and that explanations of this crisis that focus only on its structural causes thus remain incomplete. Problems in replicability and generalizability in the sciences are best explained by combining structural and psychological accounts. Correspondingly, since existing interventions to tackle scientific overgeneralizations focus primarily on structural causes, they should be supplemented with updated journal reporting guidelines and cognitive debiasing (e.g., raising awareness of generalization bias and using checklists). With targeted strategies combining structural and psychological interventions, we may significantly reduce scientific overgeneralizations and generalization bias. More experimental research is needed that recruits scientists as participants in order to evaluate the extent of generalization bias across the sciences.

Acknowledgments

For helpful feedback on earlier versions of this paper, we would like to thank the participants of the 2021 Helsinki Workshop on Reactivity in the Human Sciences as well as Stephan Guttinger, Dorothy Bishop, Olivier Lemeire, Dan Kelly, and Nikolaj Nottelmann.

Funding information

Uwe Peters and Oliver Braganza have no funding to declare. Alexander Krauss was funded by the Ministry of Science and Innovation of the Government of Spain (grant RYC2020-029424-I).

[15] For this point we are grateful to an anonymous reviewer.

References

Achinstein, P. (2010). *Evidence, Explanation, and Realism.* Oxford: Oxford University Press.

Amsel, E., Klaczynski, P. A., Johnston, A., Bench, S., Close, J., Sadler, E., & Walker, R. (2008). A dual-process account of the development of scientific reasoning: The nature and development of metacognitive intercession skills. *Cognitive Development, 23*(4), 452–471.

Baker M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604), 452–454.

Banerjee, A., & Chaudhury, S. (2010). Statistics without tears: Populations and samples. *Industrial psychiatry journal*, *19*(1), 60–65.

Bishop, D. V. (2020). The psychology of experimental psychologists: Overcoming cognitive constraints to improve research: The 47th Sir Frederic Bartlett Lecture. *Quarterly Journal of Experimental Psychology*, 73(1), 1–19. https://doi.org/10.1177/1747021819886519

Braganza, O. (2020). A simple model suggesting economically rational sample-size choice drives irreproducibility. *PloS one*, 15(3), e0229615. https://doi.org/10.1371/journal.pone.0229615

Braganza, O. (2022). Proxyeconomics, a Theory and Model of Proxy-Based Competition and Cultural Evolution. *Royal Society Open Science* 9 (2). https://doi.org/10.1098/RSOS.211030

Brownstein, M., Kelly, D. & Madva, A. (2022). Individualism, structuralism, and climate change. *Environmental Communication*, 16(2): 269-288.

Brunswik, E. (1947). *Systematic and representative design of psychological experiments; with results in physical and social perception.* U. of California Press.

Bunge, M. (1960). The Place of Induction in Science. *Philosophy of Science*, 27, 3: 262-270.

Button, K.S., Ioannidis, J.P. a, Mokrysz, C., Nosek, B. a, Flint, J., Robinson, E.S.J., and Munafò, M.R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci. 14*, 365–376.

Chomsky N. (1959). Review of Skinner's *Verbal Behavior. Language*, 35, 26–58.

Cimpian, A., Brandone, A. C., & Gelman, S. A. (2010). Generic statements require little evidence for acceptance but have powerful implications. *Cognitive science*, *34*(8), 1452–1482.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning & Verbal Behavior*, 12(4), 335–359.

Claveau, F. & Girard, J. (2019). Generic Generalizations in Science: A Bridge to Everyday Language. *Erkenntnis*, 84 (4), 839-854.

Cohen, L.J. (1970). *The Implications of Induction*. Methuen

Cohen, L.J. (1988). *An Introduction to the Philosophy of Induction and Probability*. OUP.

Croskerry, P., Singhal, G., & Mamede, S. (2013). Cognitive debiasing 2: impediments to and strategies for change. *BMJ quality & safety*, *22 Suppl 2*(Suppl 2), ii65–ii72. https://doi.org/10.1136/bmjqs-2012-001713

Cummins, R. (2000). 'How does it work?' versus 'What are the laws?': Two conceptions of psychological explanation. In F. C. Keil & R. A. Wilson (Eds.), *Explanation and cognition* (pp. 117–144). The MIT Press.

Daniel, M., Carney, M., Khandelwal, S., Merritt, C., Cole, M., Malone, M., Hemphill, R. R., Peterson, W., Burkhardt, J., Hopson, L., & Santen, S. A. (2017). Cognitive Debiasing Strategies: A Faculty Development Workshop for Clinical Teachers in Emergency Medicine. *MedEdPORTAL: the journal of teaching and learning resources*, *13*, 10646. https://doi.org/10.15766/mep_2374-8265.10646

Davidson, L. J., & Kelly, D. (2020). Minding the gap: Bias, soft structures, and the double life of social norms. *Journal of Applied Philosophy*, 37, 190–210

Dekking, M. (2005). *A Modern Introduction to Probability and Statistics*. Springer.

DeJesus, J. M., Callanan, M. A., Solis, G., & Gelman, S. A. (2019). Generic language in scientific communication. *Proceedings of the National Academy of Sciences*, 116(37), 18370–18377.

Douven, I. (2021). Abduction. *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2021/entries/abduction/>.

Dutilh Novaes, C. (2020). *Dialogical Roots of Deduction: Historical, Cognitive, and Philosophical Perspectives on Reasoning*. Cambridge University Press

Elson, M., Ferguson, C. J., Gregerson, M., Hogg, J. L., Ivory, J., Klisanin, D., Markey, P. M., Nichols, D., Siddiqui, S., & Wilson, J. (2019). Do Policy Statements on Media Effects Faithfully Represent the Science? *Advances in Methods and Practices in Psychological Science*, 12–25.

Engel, R. J., & Schutt, R. K. (2013). *The practice of research in social work* (3rd ed.). Sage Publications, Inc.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. Perspectives on Psychological Science, 8(3), 223–241. https://doi.org/10.1177/1745691612460685

Feeney, A. & Heit, E. (2007). *Inductive reasoning: Experimental, developmental, and computational approaches*. Cambridge, England: Cambridge University Press.

Fischoff, B. (1982). Debiasing. In: Kahneman, D., Slovic, P., Tversky, A., eds. *Judgment under uncertainty; heuristics and biases*. Cambridge, England: Cambridge University Press, 1982:422–44.

Fisher, A. V., Godwin, K. E., & Matlen, B. J. (2015). Development of inductive generalization with familiar categories. *Psychonomic bulletin &review*, *22*(5), 1149–1173.

Fiske, S., & Taylor, S. (2013). *Social cognition*. London: Sage.

Forscher, P. S., Mitamura, C., Dix, E. L., Cox, W., & Devine, P. G. (2017). Breaking the prejudice habit: Mechanisms, time course, and longevity. *Journal of experimental social psychology*, 72, 133–146. https://doi.org/10.1016/j.jesp.2017.04.009

Garey, M.R., & Johnson, D.S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness.* San Francisco: Freeman.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science (New York, N.Y.)*, *349*(6245), 273–278.

Giere, R. N., & Feigl, H. (Eds.). (1992). *Cognitive Models of Science* (NED-New edition, Vol. 15). University of Minnesota Press. http://www.jstor.org/stable/10.5749/j.ctttsr75

Gigerenzer, G., & Selten, R. (Eds.). (2001). *Bounded rationality: The adaptive toolbox.* The MIT Press.

Gigerenzer, G. (2018). The Bias Bias in Behavioral Economics. *Review of Behavioral Economics*, 5, 3-4, 303-336.

Guttinger, S., & Love, A. C. (2019). Characterizing scientific failure: Putting the replication crisis in context. *EMBO reports*, 20(9), e48765. https://doi.org/10.15252/embr.201948765

Haigh, M., Birch, H. A., & Pollet, T. V. (2020). Does 'Scientists Believe…' Imply ists Believe…A., & Polle'? Individual Differences in the Interpretation of Generic News Headlines. *Collabra: Psychology*, 6(1). https://doi.org/10.1525/collabra.17174

Hahn, U., & Harris, A. J. L. (2014). What does it mean to be biased: Motivated reasoning and rationality. In H. R. Brian (Ed.), *Psychology of learning and motivation* (pp. 41–102). New York: Academic Press.

Haselton, M. G., Bryant, G. A., Wilke, A., Frederick, D. A., Galperin, A., Frankenhuis, W. E., & Moore, T. (2009). Adaptive rationality: An evolutionary perspective on cognitive bias. *Social Cognition, 27*(5), 733–763. https://doi.org/10.1521/soco.2009.27.5.733

Hayes, B. K., & Heit, E. (2018). Inductive reasoning 2.0. *Wiley interdisciplinary reviews. Cognitive science*, *9*(3), e1459. https://doi.org/10.1002/wcs.1459

Heit, E., & Rotello, C. M. (2010). Relations between inductive reasoning and deductive reasoning. *Journal of experimental psychology. Learning, memory, and cognition*, *36*(3), 805–812. https://doi.org/10.1037/a0018784

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33, 61–135.

Higginson, A. D., & Munafò, M. R. (2016). Current Incentives for Scientists Lead to Underpowered Studies with Erroneous Conclusions. *PLoS biology*, 14(11), e2000995. https://doi.org/10.1371/journal.pbio.2000995

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, 23(5), 524–532. https://doi.org/10.1177/0956797611430953

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *Journal of personality and social psychology*, *103*(1), 54–69. https://doi.org/10.1037/a0028347

Kahneman, D. (2011). *Thinking, fast and slow.* New York: Farrar, Straus & Giroux.

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In *Scientific explanation*, edited by P. Kitcher and W. Salmon, 410-506. Minneapolis: University of Minnesota Press. (Minnesota Studies in the Philosophy of Science, Vol. 13)

Kukull, W. A., & Ganguli, M. (2012). Generalizability: the trees, the forest, and the low-hanging fruit. *Neurology*, 78(23), 1886–1891.

Kunda Z. (1990). The case for motivated reasoning. *Psychological bulletin*, *108*(3), 480–498.

Kurdi, B., and Banaji, M. (2021). Implicit Social Cognition: A Brief (and Gentle) Introduction. In A. S. Reber & R. Allen (Eds.), *The cognitive unconscious: The first half-century*. Oxford, UK: Oxford University Press. Retrieved from PsyArXiv. January 5. doi:10.31234/

Larrick, R. P. (2004). Debiasing. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 316–337). Blackwell Publishing. https://doi.org/10.1002/9780470752937.ch16

Leavens, D. A., Bard, K. A., & Hopkins, W. D. (2010). BIZARRE chimpanzees do not represent "the chimpanzee". *The Behavioral and brain sciences*, *33*(2-3), 100–101. https://doi.org/10.1017/S0140525X10000166

Leslie, S. J. (2017). The original sin of cognition: Fear, prejudice, and generalization. *The Journal of Philosophy*, 114(8), 393–421.

Leslie, S. J., & Gelman, S. A. (2012). Quantified statements are recalled as generics: Evidence from preschool children and adults. *Cognitive Psychology*, 64(3), 186–214.

Leslie, S.J., & Lerner, A. (2016). Generic Generalizations. The Stanford Encyclopedia of Philosophy (Winter 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/generics/>.

Lewandowsky, S., & Oberauer, K. (2020). Low replicability can support robust and efficient science. *Nature communications*, 11(1), 4109. https://doi.org/10.1038/s41467-020-17924-9

Lieder, F., & Griffiths, T. L. (2019). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *The Behavioral and brain sciences*, 43, e1. https://doi.org/10.1017/S0140525X1900061X

Little D. (1993). On the Scope and Limits of Generalizations in the Social Sciences. *Synthese*, 97, pp. 183–207.

Lockhart, J. J., & Satya-Murti, S. (2017). Diagnosing Crime and Diagnosing Disease: Bias Reduction Strategies in the Forensic and Clinical Sciences. *Journal of forensic sciences*, *62*(6), 1534–1541.

Lombrozo, T. (2013). Science: A Relationship You May Not Understand. *National Public Radio*. URL: https://www.npr.org/sections/13.7/2013/02/25/172779912/science-a-relationship-you-may-not-understand

Longino, H. E. (2002). *The Fate of Knowledge*. Princeton, NJ: Princeton University Press.

Ludolph, R., & Schulz, P. J. (2018). Debiasing health-related judgments and decision making: A systematic review. *Medical Decision Making, 38*(1), 3–13.

Mäki, U. (2001). Explanatory Unification: Double and Doubtful. *Philosophy of the Social Sciences*, 31(4), 488–506.

Mahoney, M. J. (1976). *Scientist as subject: The psychological imperative*. Cambridge, MA: Ballinger.

Martínez-Mesa, J., González-Chica, D. A., Duquia, R. P., Bonamigo, R. R., & Bastos, J. L. (2016). Sampling: how to select participants in my research study?. *Anais brasileiros de dermatologia*, *91*(3), 326–330.

Mill, J.S. (1872/1974). *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*. Eighth Edition, Toronto: University of Toronto Press.

Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing Decisions: Improved Decision Making With a Single Training Intervention. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1), 129–140. https://doi.org/10.1177/2372732215600886

Munafò, M. R., Nosek, B. A., Bishop, D., Button, K. S., Chambers, C. D., du Sert, N. P., Simonsohn, U., Wagenmakers, E. J., Ware, J. J., & Ioannidis, J. (2017). A manifesto for reproducible science. *Nature human behaviour*, 1, 0021. https://doi.org/10.1038/s41562-016-0021

Munafò, M. R., Chambers, C. D., Collins, A. M., Fortunato, L., & Macleod, M. R. (2020). Research Culture and Reproducibility. *Trends in cognitive sciences*, 24(2), 91–93. https://doi.org/10.1016/j.tics.2019.12.002

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: a call to action. *Journal of Experimental Child Psychology*, 162, 31–38.

Norton, J. (2005). A Little Survey of Induction. In Peter Achinstein (ed.) *Scientific* Evidence. Baltimore: John Hopkins University Press, 9-34.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2021). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual review of psychology*, 10.1146/annurev-psych-020821-114157. Advance online publication. https://doi.org/10.1146/annurev-psych-020821-114157

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Nature*, 349, aac4716. doi:10.1126/science.aac4716

Pelletier, F.J. (2009). *Kinds, Things, and Stuff: Mass Terms and Generics*. Oxford University Press.

Peters, U. (2020). Illegitimate Values, Confirmation Bias, and Mandevillian Cognition in Science. *The British Journal for the Philosophy of Science*. https://www.journals.uchicago.edu/doi/abs/10.1093/bjps/axy079

Peters, U. (2021). Science Communication and the Problematic Impact of Descriptive Norms. *The British Journal for the Philosophy of Science*. https://www.journals.uchicago.edu/doi/pdf/10.1086/715001

Peters, U. (2022). What Is the Function of Confirmation Bias? *Erkenntnis*. https://doi.org/10.1007/s10670-020-00252-1

Polit, D. F., & Beck, C. T. (2010). Generalization in quantitative and qualitative research: myths and strategies. *International journal of nursing studies*, 47(11), 1451–1458. https://doi.org/10.1016/j.ijnurstu.2010.06.004

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, 28(3), 369–381.

Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of Homo sapiens: Making psychological science more representative of the human population. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 115*(45), 11401–11405.

Redish, A.D., Kummerfeld, E., Morris, R.L., and Love, A.C. (2018). Opinion: Reproducibility failures are essential to scientific inquiry. Proc. Natl. Acad. Sci. U. S. A. *115*, 5042–5046.

Resnik, D. B., & Elliott, K. C. (2016). The Ethical Challenges of Socially Responsible Science. *Accountability in research*, *23*(1), 31–46. https://doi.org/10.1080/08989621.2014.1002608

Rich, P., de Haan, R., Wareham, T., & van Rooij, I. (2021). How hard is cognitive science? *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43. Retrieved from https://escholarship.org/uc/item/8cr8x1c4

Risen, J. L., Gilovich, T., & Dunning, D. (2007). One-shot illusory correlations and stereotype formation. *Personality and Social Psychology Bulletin*, 33, 1492–1502. doi:10.1177/0146167207305862

Rothbart, M., S. Fulero, C. Jenson, J. Howard, & P. Birrell. (1978). From Individual to Group Impressions: Availability Heuristics in Stereotype Formation. *Journal of Experimental Social Psychology*, 14, 237–55.

Sani, R. & Todman, J.B. (2006). *Experimental design and statistics for psychology: a first course.* Wiley InterScience.

Sellier, A. L., Scopelliti, I., & Morewedge, C. K. (2019). Debiasing Training Improves Decision Making in the Field. *Psychological science*, *30*(9), 1371–1379. https://doi.org/10.1177/0956797619861429

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin, Boston.

Shea, N., & Frith, C. D. (2016). Dual-process theories and consciousness: the case for 'Type Zero' cognition. *Neuroscience of consciousness*, 2016(1), niw005. https://doi.org/10.1093/nc/niw005

Sibbald, M., Sherbino, J., Ilgen, J. S., Zwaan, L., Blissett, S., Monteiro, S., & Norman, G. (2019). Debiasing versus knowledge retrieval checklists to reduce diagnostic error in ECG interpretation. *Advances in health sciences education : theory and practice*, *24*(3), 427–440. https://doi.org/10.1007/s10459-019-09875-8

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False positive psychology. *Psychological Science*, 22, 1359–1366.

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, 12(6), 1123–1128.

Slaney, K. L., & Tafreshi, D. (2021). Scientific generalization in psychological inquiry: A concept in need of clarification. *Qualitative Psychology*, 8(1), 82–94.

Sloman, S.A. & Lagnado, D.A. (2005). The Problem of Induction. In Holyoak, K.J., and Morrison, R.G. (eds.), *The Cambridge Handbook of Thinking and Reasoning*. Cambridge. Cambridge University Press.

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. Royal Society Open Science, 3, 160384. doi:10.1098/rsos.160384

Stanovich, K. E. (2009). What intelligence tests miss: The psychology of rational thought. Yale University Press.

Stephens, R. G., Dunn, J. C., & Hayes, B. K. (2018). Are there two processes in reasoning? The dimensionality of inductive and deductive inferences. *Psychological review*, *125*(2), 218–244.

Sutherland, S. L., Cimpian, A., Leslie, S. J., & Gelman, S. A. (2015). Memory errors reveal a bias to spontaneously generalize to categories. *Cognitive science*, *39*(5), 1021–1046.

Szucs, D., & Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. PLoS biology, 15(3), e2000797. https://doi.org/10.1371/journal.pbio.2000797

Thagard, P. (2012). *The Cognitive Science of Science: Explanation, Discovery, and Conceptual Change.* The MIT Press.

Tiokhin, L., Hackman, J., Munira, S., Jesmin, K., & Hruschka, D. (2019). Generalizability is not optional: insights from a cross-cultural study of social discounting. *Royal Society open science*, *6*(2), 181386. https://doi.org/10.1098/rsos.181386

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 76, 105–110.

Tyrer, S., & Heyman, B. (2016). Sampling in epidemiological research: issues, hazards and pitfalls. *BJPsych bulletin*, *40*(2), 57–60.

Van Rooij I. (2008). The tractable cognition thesis. *Cognitive science*, *32*(6), 939–984.

Verplanken, B. (2018). *The psychology of habit: Theory, mechanisms, change, and contexts.* Cham, Switzerland: Springer.

Welsh, M. (2018). *Bias in Science and Communication: a field guide*. Bristol, UK: IOP Press.

West, J. D., & Bergstrom, C. T. (2021). Misinformation in and about science. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(15), e1912444117. https://doi.org/10.1073/pnas.1912444117

Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science*, 34, 776–806.

Williams, J. J., Lombrozo, T., & Rehder, B. (2013). The hazards of explanation: overgeneralization in the face of exceptions. *Journal of experimental psychology. General*, *142*(4), 1006–1014.

Wu, C.M., Schulz, E., Speekenbrink, M. et al. (2018). Generalization guides human exploration in vast decision spaces. *Nat Hum Behav* 2, 915–924

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, 1-37. doi:10.1017/S0140525X20001685

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, 115(31), 7937–7942.