PRZEGLĄD PSYCHOLOGICZNY – THE REVIEW OF PSYCHOLOGY
2023, TOM 66, NR 1, 185–207
DOI: 10.31648/przegldpsychologiczny.9682

Cognitive Metascience: A New Approach to the Study of Theories

Marcin Miłkowski[1]
Institute of Philosophy and Sociology, Polish Academy of Sciences
https://orcid.org/0000-0001-7646-5742

Abstract

In light of the recent credibility crisis in psychology, this paper argues for a greater emphasis on theorizing in scientific research. Although reliable experimental evidence, preregistration, methodological rigor, and new computational frameworks for modeling are important, scientific progress also relies on properly functioning theories. However, the current understanding of the role of theorizing in psychology is lacking, which may lead to future crises. Theories should not be viewed as mere speculations or simple inductive generalizations. To address this issue, the author introduces a framework called "cognitive metascience," which studies the processes and results of evaluating scientific practice. This study should proceed both qualitatively, as in traditional science and technology studies and cognitive science, and quantitatively, by analyzing scientific discourse using language technology. By analyzing theories as cognitive artifacts that support cognitive tasks, this paper aims to shed more light on their nature. This perspective reveals that multiple distinct theories serve entirely different roles, and studying these roles, along with their epistemic vices and virtues, can provide insight into how theorizing should proceed. The author urges a change in research culture to appreciate the variety of distinct theories and to systematically advance scientific progress.

***Keywords*:** theory crisis, cognitive metascience, cognitive artifact, theoretical virtue, epistemic criteria

There are several issues that suggest a crisis of confidence in psychology and related fields. They include problems with replicability, generalizability, the cumulative nature of research, and the cohesiveness of our understanding

[1] Correspondence address: marcin.milkowski@ifispan.edu.pl.

of psychology and related fields (Baker, 2016; Boekel et al., 2015; Hughes, 2018; Ioannidis, 2005; Manninen et al., 2018; Nosek et al., 2022; Open Science Collaboration, 2015; Simmons et al., 2011; Yarkoni, 2022). While questionable research practices, fraud, inappropriate methodological choices, and poor interpretability of research due to miscommunication (Hensel, 2020; Miłkowski, 2018) are all contributing factors, it has been suggested that the unclear status and function of theory in research is at the root of the crisis, making it a *theory crisis* (Carsel et al., 2018; Hensel et al., 2022; Hughes, 2018; Irvine, 2021; Klein, 2014; Levenstein et al., 2023; Muthukrishna & Henrich, 2019; Oberauer & Lewandowsky, 2019; Smaldino, 2017; Szollosi & Donkin, 2019; Young, 2016; but see Trafimow & Earp, 2016 for a dissenting opinion). However, to evaluate this claim, a better understanding of theories is needed, including what they are and what they should do. Despite the focus on theories in the received philosophy of science, a comprehensive understanding of their roles in scientific practice is still lacking.

This paper introduces the cognitive metascience approach, aimed at offering a comprehensive understanding of theories within scientific practice. Cognitive metascience explores the processes and outcomes of evaluating scientific practices, including psychological sciences. This approach enables gaining insights into how scientific activities are perceived by practitioners, elucidating what Flis (2019) terms their "indigenous epistemology". A notable aspect of cognitive metascience is the use of both qualitative and quantitative inquiry methods, drawing from diverse disciplines, including language technology. With the advent of Big Data, it is now feasible to examine various discourse features of theoretical representations in psychology. This examination extends beyond explicit claims to include semantic relationships embedded in usage patterns. The scientific discourse on theory choice invokes normative assumptions tied to inquiry goals, indicating that both descriptive and normative insights can be derived from this research.

The subsequent section provides a more detailed presentation of the cognitive metascience approach to theories. The following section introduces considerations that suggest the presence of multiple kinds of theories in psychological science, which justifies the approach presented here. These theories serve different roles and should be evaluated based on their specific epistemic criteria for theory selection. Finally, the paper argues that a likely underlying cause of the crisis is related to the confusion between different kinds of theories and taking the theoretical background for granted. These factors can be explicitly addressed by promoting a new research culture focused on theory.

Cognitive Metascience Approach

The aim of cognitive metascience is to provide a systematic understanding and improve how scientific practice is reflected upon. Its methods draw from the philosophy of science in practice (Callebaut, 1993), cognitive science of science (Langley et al., 1987; Nersessian, 2008; Thagard & Findlay, 2012), quantitative metascience (Schooler, 2014), and digital humanities (Moretti, 2000),

including digital philosophy of science (Lean et al., 2021; Pence & Ramsey, 2018; Thagard, 1993). These methods can be used to analyze theories in various fields of psychology and yield a descriptively and eventually normatively adequate account of various kinds, functions, as well as virtues and vices of theories.

Theorizing is a cognitive practice (Callebaut, 2013; Chang, 2017) undertaken by researchers to achieve a variety of goals, including practical ones such as clinical or political aims. However, the diverse roles that theories play in scientific research can make the notion of theory somewhat elusive to many writers (Gorelick, 2011). In my view, a theory is a kind of *cognitive artifact*, a type of entity used to "maintain, display, or operate upon information in order to serve a representational function and that affect[s] human cognitive performance" (Norman, 1991, p. 11). The content of theories cannot be fully accounted for in terms of data, whether experimental or observational. Instead, theories provide a perspective on the phenomena under investigation. Multiple cognitive artifacts can form a stack that guides research practices. These artifacts may serve different functions and be tailored to the specific needs of a given domain (Miłkowski, 2022).

The representational functions of theories can be analyzed by looking at how they are implicated in distributed cognitive mechanisms (Afeltowicz & Wachowski, 2015; Giere & Moffatt, 2003; Hutchins, 1995; Osbeck & Nersessian, 2014; Zhang & Norman, 1994). Moreover, kinds of theories can be established by analyzing their distinct functions, which implies functionalism about theories. The study of function includes possible malfunctioning of theories. This implies that the aims of cognitive metascience are not only descriptive but also normative. For example, if one of the tasks of theory is to classify phenomena awaiting explanation, then a theory that does not classify the phenomena in an appropriate fashion will be deficient in that respect. In general, cognitive metascience aims to provide a normative framework for the evaluation of cognitive artifacts, including theories, and to improve the quality of scientific theorizing.

With this broad understanding of theories, we can now focus on how the credibility crisis is supposed to be related to the lack of suitable theories in psychology, which is a common complaint (Fiedler, 2017; Muthukrishna & Henrich, 2019; Oberauer & Lewandowsky, 2019). Why should unsuitable theories play any role?

One obvious answer is that theories constrain hypotheses put forward by researchers. The predictive value of a hypothesis depends to a large extent on *a priori* probability of the hypothesis' being true: if most hypotheses we test are false, then it is only to be expected that a large proportion of empirically confirmed hypotheses are also likely false. Since it is theory that should constrain the pool of seriously entertained hypotheses, the suggestion is that a theory crisis might be at the root of the replication and generalizability crisis (Hensel et al., 2022). The rational evaluation of theories may be biased if only a limited number of alternative theories is entertained (Almaatouq et al., 2022; Dellsén, 2020).

Theories may also fail to appropriately constrain predictive or explanatory hypotheses (Bird, 2021; Button et al., 2013; Fiedler, 2017; Muthukrishna & Henrich, 2019). Bird (2021), for instance, insists that the failures of replication should be expected in many fields of inquiry. However, his interpretation of replication failures seems to overlook the differences in replicability levels among relatively

similar scientific fields, and within fields, making it too coarse-grained for understanding such differences (Autzen, 2021).1

Replication may also be hindered by inappropriate theoretical accounts of observed phenomena (Eronen & Bringmann, 2021; Levenstein et al., 2023) and by confusing mere descriptions of the phenomena with their explanations (Scheel et al., 2020). Gigerenzer (1998) observed that psychologists tend to use surrogates for proper theories: one-word explanations, redescriptions, muddy dichotomies, and data fitting. In short, there are numerous suggestions as to how the crisis may be related to the improper use of theories.

Psychological research at times seems to either have too little theory or rely on theory too much. On the one hand, researchers in many fields of psychology engage almost exclusively in gathering experimental and observational data, fishing for effects, even if they do not constitute targets of explanation (Cummins, 2000; Fried, 2020; van Rooij & Baggio, 2021). On the other hand, once developed, an attractive enough theory may seem to obviate the need for any observations. Indeed, many existing theories seem to be irrefutable because they can accommodate almost any empirical finding (Frankenhuis et al., 2023; Miłkowski & Litwin, 2022; Roberts & Pashler, 2000; Szollosi & Donkin, 2019). Alternatively, many defenders of a theory focus on its confirmation, ignoring the need to compare it to competitors (Greenwald et al., 1986). Theories are also rarely if ever rendered obsolete (Ferguson & Heene, 2012).

The lack of suitable theory may lead to a range of consequences beyond low replicability and generalizability. If psychology proceeds in an atheoretical manner, it is likely to offer ad hoc explanations and fail to be cumulative (Erdin, 2021; Newell, 1973). According to many, the core issue is the lack of theoretical unity in psychology, which is separated into distinct and inconsistent approaches or paradigms (Hughes, 2018). One possible risk of the lack of theoretical unity, as argued, is that the disunified research is fragmented and disintegrated, providing no systematic interconnections between phenomena (Bower, 1993; Goertzen, 2008; Staats, 1986; Young, 2016; but see Dale et al., 2009; Matthews, 2020; Zittoun et al., 2009; for a review, see Gaj, 2016).

Moreover, without a clearly specified theory, one cannot tell which results depend on core claims and which on auxiliary assumptions (Cooper & Shallice, 1995; Miłkowski et al., 2018). And even when a psychological theory generates predictions, they tend to be vague (Fried, 2020; Meehl, 1967). This theoretical vagueness may actually serve strategical purposes of researchers, while remaining detrimental to scientific progress (Frankenhuis et al., 2023). But this leaves us with the question of what makes theories suitable and for what purposes.

Fortunately, to evaluate theories as cognitive artifacts, one may rely on the entrenched notion of theoretical virtue, which has become influential in methodology since the 1970s. This notion can be used to clarify connections between particular features of theories and possible causes of the crisis. It was introduced by Thomas Kuhn (1977), who identified five theoretical virtues for scientific investigation in general: empirical accuracy, consistency, scope, simplicity, and fruitfulness. Interestingly, systematicity, although arguably a significant characteristic of science (Hoyningen-Huene, 2013; Rescher, 1979), as well as the capacity to

design replicable experiments, is missing not only from Kuhn's list but also from subsequent research on theoretical virtues (Keas, 2018; Longino, 1996; McMullin, 2010; Schindler, 2018). However, even when brain and mind researchers do recognize the significance of theoretical virtues (e.g., Muthukrishna & Henrich, 2019), they rarely reflect on the specific functions and properties of theories.

For all the importance of the virtues identified by Kuhn, it should be noted that methodological standards may vary from one specific research field to another (Laudan, 1984). For example, in computational modelling, tractability (Dror & Gallogly, 1999; Frixione, 2001; van Rooij, 2008; van Rooij & Baggio, 2021) is such a virtue. Therapeutic translatability to clinical practice is another example from clinical psychology. There need not be a single list of necessary and sufficient properties that any satisfactory theory should display; these properties depend on the functions of theories.

However, Kuhn's insistence that virtues are similar to other values goes too far. While at least some values can be freely chosen, this is not the case with epistemic virtues (Norton, 2021). Instead, as Norton insists, these are epistemic criteria of theory choice. In the framework presented here, these criteria are derived from the functionality of various kinds of theories as cognitive artifacts, and they cannot be imposed from the armchair without disrupting their cognitive roles. That means that virtues and vices should be evaluated predominantly by focusing on functions of theories. The task of the next section is therefore to propose a preliminary functional taxonomy of theories and associate them with their particular virtues and vices. A detailed empirical study of particular kinds of theories and their virtues is, unfortunately, beyond the scope of the current paper, but the aim of the next section is to legitimize the cognitive metascience approach to theory issues.

Kinds, Functions, and Virtues of Theories

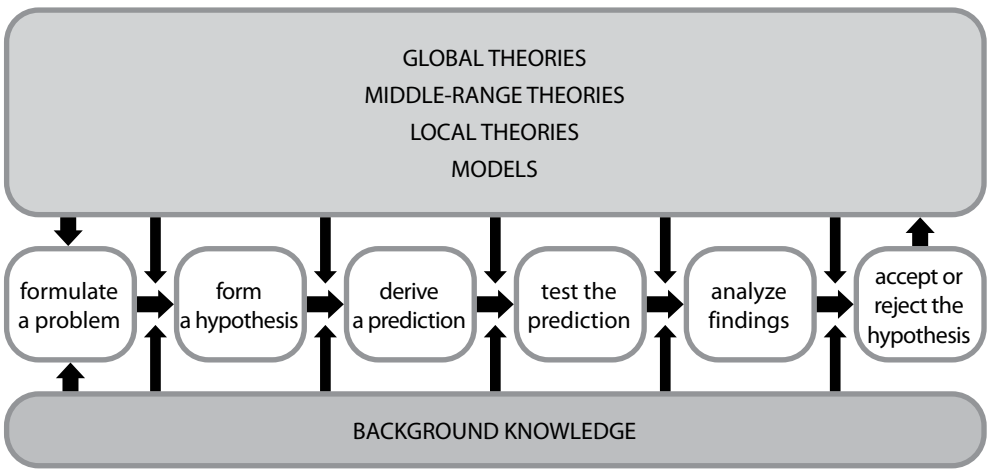
Theoretical assumptions that go beyond experimental or observational data are pervasive throughout the research process. In this section, I propose a preliminary taxonomy of theories based on previous insights from methodological reflections, enriched by scientific discourse analysis. The taxonomy is preliminary and aims to justify further systematic inquiry into their functionality from the perspective of cognitive metascience. This paper is the first step towards gaining a new understanding of the different kinds of theories involved not only in psychological research but also in the sciences of the brain and mind, broadly considered. The task here is to argue for the plausibility of the claim that there are multiple kinds of theories, each with unique assessment standards.

To illustrate my claims in more detail, I will distinguish several kinds of theories, which can be related to a simplified course of confirmatory research (see Figure 1). The main point of this section is that there is more than one kind of theory. While there are plausibly other kinds of theories (e.g., local theories vs. middle-range theories, normative vs. descriptive theories), I will focus merely

on the course of theorizing over several stages of inquiry. In reality, of course, the progression need not be chronological (e.g., analytical tools may obviously exist earlier), but it helps my presentation purposes.

Figure 1

All stages of research are theoretically informed



Let us then start with the elementary step in any experimental or observational study: regarding something as a problem. This already requires some theoretical commitment. Notably, the targets of eventual explanation are neither raw data (see also Aronova et al., 2017; Gitelman, 2013), nor data themselves. It has been argued that what is explained are phenomena (Bogen & Woodward, 1988) or models of data (Suppes, 1962). While these proposals are somewhat different, the core claim is the same: explanations do not explain data as such; they explain theory-laden data about phenomena. The same goes for theory-driven predictions: one is rarely if ever interested in predicting, for example, measurement errors (as present in experimental data). Theories of data play a predominantly descriptive role, by delineating targets of our inquiry. While they include theoretical terms, or concepts (such as "grammatical competence", which is itself unobservable, because the capacity to parse an indefinite number of sentences in a language is not observable), these theories also indicate essential facts that characterize referents of these concepts, in particular regularities that pertain to them. In contrast, mere collections of specimens or instances can be fully observable, and they carry no theoretical commitment.

However, *theories of data* (or phenomena) are not explanatory or predictive themselves – or at least this is not their primary purpose. In the case of complex theories of data, one can speak of *classificatory theories* (Leonelli, 2016) such as found in diagnostic manuals in psychiatry (American Psychiatric Association, 1980), or cognitive ontologies (Poldrack et al., 2011). This kind of theory determines also how one delineates and identifies mechanisms responsible for

the phenomena (Craver, 2007, 2009). Plausibly, lack of proper theories of data is one of the causes of the ongoing crisis (Eronen & Bringmann, 2021; Levenstein et al., 2023). One of the hypothesized causes of the credibility crisis in psychology is also confusing theories of data with explanatory theories (Scheel et al., 2020).

At the next stage, one should be able to form a hypothesis. This is the task of theoretical considerations of *hypothesis-forming*, which may be guided by heuristics associated with a given theory or research tradition. In cognitive science, research traditions frequently employ fallible heuristics to generate hypotheses (Fiedler, 1991; Miłkowski, 2019). The heuristics used to generate hypotheses may also arise from particular methods or tools adopted by researchers (Gigerenzer, 1991, 1992; Greenwald, 2012; Nickles, 2018). These purely *heuristic theories* can function in addition to proper *explanatory* or *predictive* theories.

While philosophers of science and researchers are familiar with examples of theories that serve solely a predictive function, they often presuppose that the job of theories is to explain. Although merely predictive theories may fall short of providing rich explanations (such as signal detection theories), they still seem justifiably called "theories" because they go beyond the given data.

For instance, it would be hasty to dismiss all signal detection theories as a predictive tool in psychology, since the primary purpose of such theories is to predict unseen observable data, which can be useful for various other purposes, even if it fails to provide a satisfactory explanation for them. Additionally, assuming that merely predicting unseen data indicates explanatory power of a theory is a significant error (Shmueli, 2010).

Psychology has yet to fully recognize the potential of predictive theories, which are evaluated using methods such as cross-validation rather than traditional significance testing. Such theories may offer more generalizable inferences than local explanatory theories (Yarkoni & Westfall, 2017). Notably, scientific researchers differentiate between "predictive understanding" and "explanatory understanding" in practice. This can be evidenced using language technology, by consulting the DOAJ Corpus of open access journals, available through the advanced Sketch-Engine web corpus software (Kilgarriff et al., 2014).[2] However, even philosophers who emphasize the value of prediction, like Broadbent (e.g., Broadbent, 2018) sometimes mistakenly conflate understanding with explanatory understanding. Cognitive metascience, in contrast, recognizes that these two can diverge. Nonetheless, while some explanatory theories can be "converted" into predictive ones (Shmueli & Koppius, 2011), this process is complex and highlights the different roles that predictive and explanatory theories play in research.

Sometimes general theories do not provide testable hypotheses directly, but motivate or constrain models of the phenomena instead. Although models are relatively independent of theories, they act as mediators between theory and reality (Morgan & Morrison, 1999). While the relation between a theory and its model may be more or less tenuous, general theories are reasonably expected to provide systematic motivation for models. These are *model-motivating* theories.

[2] The source data is available for download at https://osf.io/df3ru/ [as of April 18, 2023].

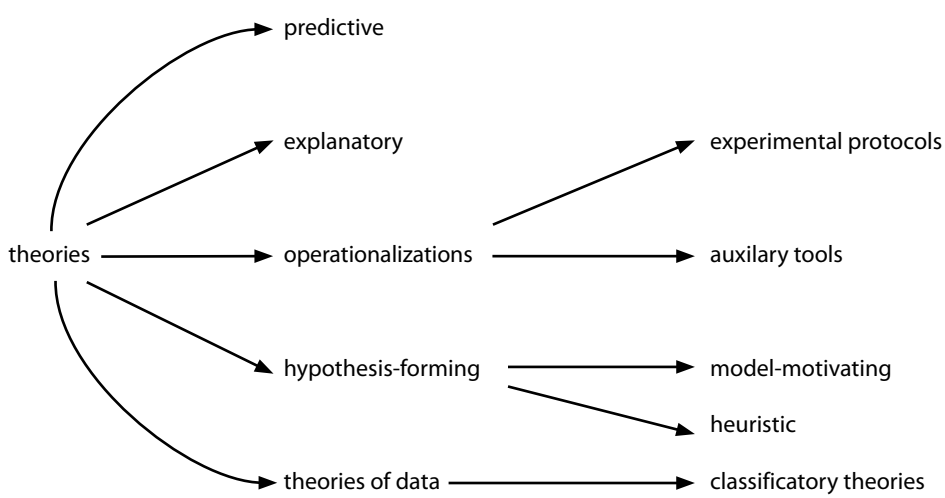
The hypothesis or model may require further operationalization to provide testable predictions. These operationalizations are not necessarily parts of the overarching explanatory theory but may become entrenched in research. E.g., one may rely on eye fixation in babies to provide insight into their attention. Eye movements and fixation are then used for designing experimental conditions. Moreover, at the next stage of the research process, additional insights are required to design the experiment. What comes into play here are not only the usual experimental protocols, which vary from one field to another (Sullivan, 2009), but also auxiliary tools (e.g., to make statistically robust inferences). While operationalizations do not provide novel predictions, they are sufficiently important for differences between them not to be overlooked when attempting research replication. Operationalizations are supposed to render the phenomena, described by theories of data, open to empirical inquiry (observational or experimental). An operationalization is correct, or has high construct validity, to the extent to which the observable variable selected by the researcher reflects the presence of the appropriate theoretical construct. Given the fact that most phenomena studied by psychology are not directly observable, the quality of operationalizations has immense significance for the validity and reliability of research (Hughes, 2018).

Finally, after the data have been collected and preprocessed, statistics and other analytical tools are used along with the original theory to establish whether the evidence confirms or disconfirms the null hypothesis. These auxiliary or analytic tools can have a dramatic impact on the results of empirical research. Recent studies indicate that the same dataset can be analyzed in divergent ways by various research teams, which contributes to the theoretical flexibility in mind and brain sciences (Botvinik-Nezer et al., 2020; Silberzahn et al., 2018).

Figure 2 provides a diagram depicting the preliminary taxonomy described above.

Figure 2

Preliminary taxonomy of kinds of theories in confirmatory research



Admittedly, the taxonomy provided here is merely a sketch of what is at stake in psychological research. For example, one could distinguish explanatory and predictive theories that aim at describing the group averages from ones that focus on individuals, and there are also normative theories, as opposed to descriptive ones. My aim here is to motivate further systematic empirical research into theoretical representations already inherent in scientific practice. At the same time, it should be quite clear that the taxonomy illustrates a variety of theoretical representations in science: there is nothing in experimental data about attention in babies, for example, that would force a researcher to use Bayesian statistics as her auxiliary tool. Some of the theoretical choices, obviously, can be motivated by pointing that these lead to robust predictions (e.g., this is how eye fixation times can be justified as operationalizing attention) but they do go beyond the observable data.

Having established that there is some unity in all this diversity–all these cognitive artifacts transcend in their contents what is available in the observational or experimental data–it is important to appreciate their diversity in terms of their methodological assessment. Not all kinds of theories must share the same set of epistemic criteria. In other words, one should not mechanically impose Kuhn's (1977) five theoretical virtues on all of them. Let me start from theories of data: these need not be fruitful, i.e., systematically lead to new discoveries, for example by providing research questions and discovery heuristics (Ivani, 2019). Even classificatory theories need not be tasked with providing us with research questions: the task of a diagnostic manual of mental pathologies, for example, is to help clinicians in (statistically reliable) diagnoses rather than in the inquiry into their causes. Of course, it could be a welcome *additional* feature of such a classificatory theory, but it is simply not necessary for its proper functioning in classifying only certain symptoms as indicating particular pathologies.

Interestingly, it is often assumed that hypothesis-forming theories may be much more fallible than other kinds of theories, i.e., it need not be robustly empirically accurate. In particular, it is frequently assumed that heuristic theories are fallible. One such example can be easily adduced: embodied cognition research tradition in cognitive science provides an experimental heuristic that suggests that at least some variance in performance in cognitive tasks may be explained by appealing to bodily factors–such as body morphology or physiological state–even if this heuristic could fail in the majority of cases (Miłkowski, 2019). It is understood that this is still a valuable heuristic because it could lead to more empirical discoveries: Heuristic theories must be fruitful, in contrast to classificatory theories. Model-motivating theories may fail to be accurate in another fashion: They could provide just a general scheme for building models. It is arguable, for instance, that predictive processing theories of cognition are not detailed enough to provide specific predictions, but they are sufficient to provide motivation for building computational models for a vast number of experimental phenomena (Litwin & Miłkowski, 2020).

Suppose we have classified a phenomenon, formed a hypothesis or even built a computational model. Time for designing an experiment: should one's experimental operationalization of the phenomenon be externally consistent with

the previous one? Not at all. Operationalizations often vary dramatically and improve over time. This means that while they should be consistent with other theoretical assumptions used in the same experimental setup (and the overall context of inquiry), they need not be particularly consistent with what has been claimed before. Note that classificatory theories usually vary in a much less dramatic fashion: changes in diagnostic manuals usually take years and a committee, not a single journal paper, as these are much more entrenched (Eronen & Bringmann, 2021). Similarly, explanatory theories usually must retain consistency with their overall research tradition much more than particular operationalizations. Improving an experimental protocol seems the most flexible feature of experimental research.

Now, let us then try to assess the explanatory power of the hypothesis formed by building a full-blown explanatory theory. Should we aim for the broadest scope? Not really. Explanatory theories need not be extremely general. For example, so-called micro-theories, or theories of individual subject performance in a task, may not generalize across multiple subjects whose learning histories vary (Newell & Simon, 1972). Therefore, their scope of breadth need not be huge; they are not designed to generalize over subjects with diverging learning histories. Moreover, explanatory theories need not be predictive (this feature is not included by Kuhn for an obvious reason: he seems to presuppose that theories proper are explanatory theories), and predictive theories need not be explanatory.

It should be now clear why confusing various kinds of theories may be one of the underlying causes of the theory crisis. This is because of two significant considerations. First, it's essential to assess different types of theories using different sets of criteria. Second, the failure to appreciate that a complete account of a psychological phenomenon requires developing several, somewhat independent, theoretical representations may lead to incomplete results reporting, which, in turn, can contribute to lower replicability. This is because theoretical descriptions of the phenomena at hand are crucial in any attempt to replicate the study in a different experimental setting: these descriptions determine which factors should be controlled when devising the experimental conditions and the control condition. Importantly, even in relatively close research fields, typical experimental protocols for the study of the similar phenomena may vary significantly, which can defy attempts to integrate their experimental results (Sullivan, 2009).

To sum up, the reason why we should care about distinctions among various kinds of theories is that they should not be assessed using the same standards. They serve their specific functions, and keeping several theories separate may actually contribute to more reproducible and replicable research. It can be also fruitful for a healthy research culture.

Towards a Culture of Theory

The stack of theories inherent in any research study usually entails a tangled web of scientific representations that drives the inquiry process. While

the theories involved can be dissociated, they are usually intertwined. For example, distinct theoretical hypotheses may be used to explain the same phenomenon, while sharing the same theory of data.

Take the spectacular example of the reported replication failure of ego depletion (Hagger et al., 2016). Before the failure, almost 600 studies were published that reported the ego depletion effect. All of these shared a basic theory of data: one kind of mental fatigue was believed to be induced experimentally, which was later believed to diminish the subjects' self-control ability in subsequent tasks. However, the multilab replication study reported a null effect, which should weaken the belief in the accuracy of the data theory (in particular because the experimental protocol was vetted by the original researchers). Null effects are typically assumed to require no explanation, and various distinct hypotheses of the nature of the "mental energy" (or glucose levels) behind the effect were thereby undermined. That means that the explanatory hypotheses for ego depletion, even if posed independently of the theory of data, make sense only insofar as the theory of data stands scrutiny. Arguably, this could lead to the demise of the original theory of ego depletion and the popularity of the experimental protocol (Vadillo, 2019).

However, in this case, the null effect seems to be in need of an explanation. In particular, we should be able to see whether the convergent results of 600 studies can be explained as mere false positives, among a huge number of other similar results in psychology (Simmons et al., 2011), and whether the lack of publication of negative results is only due to the file-drawer effect (Rosenthal, 1979), whereby null effects are rarely ever published.

Nonetheless, even a multilab replication failure can be contested, for example by performing yet another multilab replication with a different experimental protocol (Dang, 2016; Vohs et al., 2021), and the main proponent of the studies over ego depletion, Roy Baumeister, boldly claims that it is the best replicated phenomenon in all social psychology (Baumeister & Tice, 2022). Indeed, the failure could be due to the experimental task that supposedly did not match the protocol used in original hundreds of studies. In this case, the original proponent claims that the phenomenon is real, but the experimental intervention should be performed in a way that matches his studies. In other words, theory of data would be still accurate, while experimental protocol inappropriate. The critics, however, claim that inducing a certain kind of mental fatigue over very short time spans may as well have little if any effect on mental processing. In other words, they are skeptical of whether the measurement procedures used in the protocol display any construct validity.

This short example shows that for any replication failure, several theoretical representations could be blamed. These include theories of data, where the phenomenon could be an artifact of questionable research practices; operationalizations, including experimental protocols with their particular choice of dependent and independent variables; and auxiliary statistical tools, which are frequently the focus of subsequent analyses. Finally, there are explanatory theories that guide the experiment design, providing explanatory factors along with possible new experimental predictions to be tested in the future. All of these are partially

or wholly independent, even if used together, and the failure to replicate may be insufficiently diagnostic to indicate which of these is the main culprit.

Replication efforts may falter due to fraud or other questionable research practices, inappropriate methodological choices, exacerbated by publication bias (which includes file-drawer effects on negative results), or even as a result of underspecified designs in the original study publications (e.g., frequently missing self-report scales commonly employed in many psychology studies). However, setting aside these factors, theory-related considerations emerge as particularly plausible candidates to scrutinize. Put differently, representations tied to theory play a pivotal role in scientific inquiry, and their validity must be assessed in instances of replication failures.

This suggests that explicit theory development is indeed indispensable to deal with the current credibility crisis, but it should be more specifically understood. Researchers must explicitly develop several theories, including theories of data and valid operationalizations, to make their results robust. Some of these theories remain thin and underspecified nowadays, possibly because general psychological principles are considered too trivial to be interesting (Hensel et al., 2022). However, systematic development of multiple theories can only help in taking the first steps towards dissolving the credibility crisis.

The next step is clearly to develop and discuss the virtues and vices of theoretical representations. This is already happening in the current debate, in which procedures are proposed not only to minimize the possible impacts of questionable practices (such as unreported rejection of data inconsistent with initial hypotheses as outliers) but also to discuss the soundness of methods. However, it is not just the methods that should be sound; theories should be sound as well. A major problem with the modern version of the ego-depletion explanatory theory is that it links glucose levels with the effect, and glucose levels cannot possibly vary dramatically over relatively short time spans (Vadillo et al., 2016). In other words, the theory seems outright inconsistent with (much more empirically robust) physiological theories of glucose metabolism. Nonetheless, without glucose levels as the resource to be depleted during self-control tasks, the ego-depletion theory becomes uninformative and vague.

For explanatory theories, there is already a fairly explicit discussion of commonly assumed virtues and corresponding vices (Keas, 2018), even if it misses some criteria used in practice (such as systematicity). We are still missing this kind of discussion for other kinds of theoretical representations, including experimental protocols. However, with the help of cognitive metascience approach driven by language technology, one can identify that there are some predicate adjectives, which ascribe properties to experimental protocols in the vast body of literature (again from the DOAJ corpus). These are “identical”, “suitable”, “specific”, “similar”, “consistent”, and “necessary”. There are uses of “identical” that mark a cognitive virtue: when reproducing an experimental result, one should use the identical experimental protocol (this is the necessary condition of reproducibility); other uses may simply state that for multiple studies reported in the paper or a series of papers, the same (identical) protocol was used, which is unrelated to virtues. While there is insufficient space in this paper to focus on

details, this kind of evidence can be helpful in providing more insight for improving current research practices. In particular, not only the validity of the protocol should be checked, but also its interoperability and openness.

The intersubjective validity of the experimental protocol is related to one of the recurring problems in psychology, that of the intersubjective validity of its theories of data. For example, there are recurring worries regarding the interrater reliability of diagnostic criteria for mental disorders. One revolutionary aspect of the DSM-III was its stress on reliability, likely due to the perceived problems with the flexibility of diagnoses (Kawa & Giordano, 2012; Wilson, 1993). Some of the apparent crisis with diagnoses was overblown, as the infamous study by Rosenhan (1973), which claimed that diagnoses in mental institutions were entirely arbitrary and premature, turned out to be a case of scientific fraud (Cahalan, 2019; Scull, 2023). Making diagnoses reliable is one case where undesired theoretical flexibility is being removed. For the purposes of this paper, it is sufficient to point out that DSM-III was itself a response to the perceived crisis, and that the explicit effort to make it reliable across multiple diagnosticians is indeed a laudable feature of this nosological theory.

One way to evaluate the DSM or any theory of data is to determine whether it enables successful measurement, in addition to categorization. Successful measurement has at least two key features: it is precise and convergent (Isaac, 2019). Precision is demonstrated when measurement procedures consistently produce the same results (within similar error bounds) when performed repeatedly (in the DOAJ corpus, one can find that this property is ascribed to measurement using such terms as “reproducible” and “repeatable”, which makes it clear how this can be linked to the reproducibility and replicability of research). In the case of diagnoses, this means that they should be precise over time and across multiple diagnosticians. A significant increase in precision, in terms of interrater reliability, has been observed in clinical practice when transitioning from DSM-I or DSM-II to DSM-III (Di Nardo et al., 1983). However, achieving convergence, which involves obtaining the same result using different procedures, is more challenging because divergent theoretical commitments inherent in these procedures can influence diagnoses. Achieving convergence could support a realistic or objective approach to diagnoses and would be sufficient to demonstrate construct validity. Unfortunately, this is still but a pipe dream.

Moreover, DSM-III, like many other qualitative judgments, can be understood in measurement terms in a stretched sense, as relying on nominal or ordinal scales that do not have all the features of metric spaces used for measurement in other fields (for a more nuanced analysis of how DSM could be viewed from a psychometric point of view, see, e.g., (Borsboom, 2008)). Nonetheless, while it is unlikely that all psychology can be made similar to psychophysics in its reliance on metric spaces (sets of elements with a distance metric defined over them), some progress is still possible. It is arguable, for example, that recent attempts to provide an alternative nosology based on networks of symptoms (Borsboom et al., 2018; Borsboom & Cramer, 2013), are a step in this direction.

While no diagnostic manual in psychopathology can take pride in being fully satisfactory, the perspective taken in this paper allows us to say that reliability

is a necessary, albeit insufficient, epistemic criterion of choice for theories of data. It removes one source of the crisis, which is flexibility, both of measurement procedures and theoretical descriptions. However, the road to a fully successful theory is not automatically open.

This is not only because measurement problems are not the sole reason for the current crisis. In fact, while progress in measurement can contribute to theoretical progress, theoretical progress can outstrip our measurement capacities (Bringmann & Eronen, 2016). This was the case of Galileo, who could not measure the time duration reliably (Koyré, 1953). Arguably, computational simulation in psychology provides examples of precise theories (Anderson, 2007; Marr, 1982), whose full empirical evaluation is immensely difficult, and we still lack appropriately fine-grained neuroscientific evidence to assess some of their claims. In fact, theoretical progress can rely simply on a better conceptualization of the phenomena to be explained. This was the case of Chomsky's (1959) criticism of Skinner's approach to language. By pointing out that language users are capable of understanding and producing indefinitely many grammatically correct statements (which is dubbed "productivity"), he provided a plausible theory of observable linguistic data regarding our behavior. The main argument was that Skinner's theory could not account for this productivity. Even if some parts of his criticism could be rebutted (MacCorquodale, 1970), the theoretical claim that human language use displays productivity remains one of the entrenched assumptions for further research.

Making theory development more prominent in psychological research requires both small and larger steps to take root. Here, I have only pointed out progress in theories of data, rather than explanatory or predictive theories. However, it is essential to achieve progress in all kinds of theoretical insights. This requires joint efforts, but the lack of cumulative work on theories has led to theoretical fragmentation and a lack of shared efforts on theory development. Theories are treated akin to somebody else's toothbrushes (Mischel, 2008). To address this, we need collaborative efforts, including adversarial collaboration, on theoretical issues (Cowan et al., 2020; Del Pin et al., 2021), in addition to systematic reviews. We need to let a thousand flowers bloom, but theories should be pruned when they become obsolete. For this, we need more explicit focus on theory assessment.

Conclusions

The purpose of this paper was to justify a new approach to the study of science: cognitive metascience, particularly in times of growing perception of the severity of the crisis in psychology. By looking at theories as theoretical artifacts whose contents cannot be accounted for only in terms of experimental or observational data, it opens a new perspective on theoretical representations whose roles are not only for explanation but also for describing phenomena, predicting, operationalizing, and forming new hypotheses. This broad understanding of theories

should be complemented by the appropriate application of the distinct normative criteria used to evaluate these representations in scientific practice.

Without appropriately developed theories, we cannot hope for putting all together in a successful theoretical proposal for psychological phenomena, and deal with multiple ailments that were recently discussed in metascientific debates. The credibility crisis requires a complex answer, and one part of it is more stress on theories and their assessment. For this, the change of research culture is indispensable. This change should lead to the development of shared and more stringent epistemic criteria for theory assessment.

Due to space constraints, I have only briefly touched upon the distinct roles and associated virtues (and vices) of different types of theories. Further research will focus on specific types of theoretical representations, both qualitatively and quantitatively, to gain a more nuanced understanding of theories. Cognitive metascience aims to provide general insights, drawing inspiration from the philosophy of science, but deviating from the practices of science and technology studies and cognitive science of science, which typically focus on individual labs or historical cases. By leveraging Big Data, digital humanities methodologies enable "distant reading" (Moretti, 2000) providing a way to generalize beyond individual case studies (Piper, 2020). While cognitive scientists have thoroughly studied the functioning of cognitive artifacts in various cases, we must be able to systematically and comprehensively study the functional types of theories to yield general normative advice. This is humanly impossible in its generality without the aid of language technologies. Such studies, however, will be a topic for our future work.

Data availability

The source concordances and collocations from the DOAJ corpus, generated from SketchEngine (in the CVS format) are available for download at https://osf.io/df3ru/.

References

Afeltowicz, Ł., & Wachowski, W. (2015). How Far we Can Go Without Looking Under the Skin: The Bounds of Cognitive Science. *Studies in Logic, Grammar and Rhetoric*, *40*(1), 91–109. https://doi.org/10.1515/slgr-2015-0005

Almaatouq, A., Griffiths, T. L., Suchow, J. W., Whiting, M. E., Evans, J., & Watts, D. J. (2022). Beyond Playing 20 Questions with Nature: Integrative Experiment Design in the Social and Behavioral Sciences. *Behavioral and Brain Sciences*, 1–55. https://doi.org/10.1017/S0140525X22002874

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). American Psychiatric Association.

Anderson, J. R. (2007). *How Can the Mind Occur in the Physical Universe?* Oxford University Press.

Aronova, E., Oertzen, C. von, & Sepkoski, D. (Eds.). (2017). *Data histories*. University of Chicago Press.

Autzen, B. (2021). Is the replication crisis a base-rate fallacy? *Theoretical Medicine and Bioethics*, *42*(5), 233–243. https://doi.org/10.1007/s11017-022-09561-8

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature News*, *533*(7604), 452. https://doi.org/10.1038/533452a

Baumeister, R. F., & Tice, D. M. (2022). Ego Depletion is the Best Replicated Finding in All of Social Psychology. *Scholarly Journal of Psychology and Behavioral Sciences*, *6*(2), 686–688. https://doi.org/10.32474/SJPBS.2021.06.000234

Bird, A. (2021). Understanding the Replication Crisis as a Base Rate Fallacy. *The British Journal for the Philosophy of Science*, *72*(4), 965–993. https://doi.org/10.1093/bjps/axy051

Boekel, W., Wagenmakers, E.-J., Belay, L., Verhagen, J., Brown, S., & Forstmann, B. U. (2015). A purely confirmatory replication study of structural brain-behavior correlations. *Cortex*, *66*, 115–133. https://doi.org/10.1016/j.cortex.2014.11.019

Bogen, J., & Woodward, J. (1988). Saving the Phenomena. *The Philosophical Review*, *97*(3), 303. https://doi.org/10.2307/2185445

Borsboom, D. (2008). Psychometric perspectives on diagnostic systems. *Journal of Clinical Psychology*, *64*(9), 1089–1108. https://doi.org/10.1002/jclp.20503

Borsboom, D., Cramer, A., & Kalis, A. (2018). Brain disorders? Not really… Why network structures block reductionism in psychopathology research. *Behavioral and Brain Sciences*, 1–54. https://doi.org/10.1017/S0140525X17002266

Borsboom, D., & Cramer, A. O. J. (2013). Network Analysis: An Integrative Approach to the Structure of Psychopathology. *Annual Review of Clinical Psychology*, *9*(1), 91–121. https://doi.org/10.1146/annurev-clinpsy-050212-185608

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., … Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9

Bower, G. H. (1993). The fragmentation of psychology? *American Psychologist*, *48*(8), 905–907. (1994-00003-001). https://doi.org/10.1037/0003-066X.48.8.905

Bringmann, L. F., & Eronen, M. Ilkka. (2016). Heating up the measurement debate: What psychologists can learn from the history of physics. *Theory & Psychology*, *26*(1), 27–43. https://doi.org/10.1177/0959354315617253

Broadbent, A. (2018). Prediction, Understanding, and Medicine. *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine*, *43*(3), 289–305. https://doi.org/10.1093/jmp/jhy003

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365. https://doi.org/10.1038/nrn3475

Cahalan, S. (2019). *The Great Pretender*. Grand Central Publishers.

Callebaut, W. (1993). *Taking the naturalistic turn or how real philosophy of science is done*. University of Chicago Press.

Callebaut, W. (2013). Naturalizing Theorizing: Beyond a Theory of Biological Theories. *Biological Theory*, *7*(4), 413–429. https://doi.org/10.1007/s13752-013-0122-2

Carsel, T., Demos, A. P., & Motyl, M. (2018). Strong scientific theorizing is needed to improve replicability in psychological science. *Behavioral and Brain Sciences*, *41*, e123. https://doi.org/10.1017/S0140525X1800078X

Chang, H. (2017). VI – Operational Coherence as the Source of Truth. *Proceedings of the Aristotelian Society*, *117*(2), 103–122. https://doi.org/10.1093/arisoc/aox004

Chomsky, N. (1959). Review of Verbal Behavior by B. F. Skinner. *Language*, *35*(1), 26–58.

Cooper, R. P., & Shallice, T. (1995). Soar and the case for unified theories of cognition. *Cognition*, *55*(2), 115–149. https://doi.org/10.1016/0010-0277(94)00644-Z

Cowan, N., Belletier, C., Doherty, J. M., Jaroslawska, A. J., Rhodes, S., Forsberg, A., … Logie, R. H. (2020). How Do Scientific Views Change? Notes From an Extended Adversarial Collaboration. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *15*(4), 1011–1025. https://doi.org/10.1177/1745691620906415

Craver, C. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press.

Craver, C. F. (2009). Mechanisms and natural kinds. *Philosophical Psychology*, *22*(5), 575–594. https://doi.org/10.1080/09515080903238930

Cummins, R. (2000). "How does it work" versus "what are the laws?": Two conceptions of psychological explanation. In F. Keil & R. A. Wilson (Eds.), *Explanation and Cognition* (pp. 117–145). MIT Press.

Dale, R., Dietrich, E., & Chemero, A. (2009). Explanatory Pluralism in Cognitive Science. *Cognitive Science*, *33*(5), 739–742. https://doi.org/10.1111/j.1551-6709.2009.01042.x

Dang, J. (2016). Commentary: A Multilab Preregistered Replication of the Ego-Depletion Effect. *Frontiers in Psychology*, *7*, 1155. https://doi.org/10.3389/fpsyg.2016.01155

Del Pin, S. H., Skóra, Z., Sandberg, K., Overgaard, M., & Wierzchoń, M. (2021). Comparing theories of consciousness: Why it matters and how to do it. *Neuroscience of Consciousness*, *2021*(2), niab019. https://doi.org/10.1093/nc/niab019

Dellsén, F. (2020). The epistemic impact of theorizing: Generation bias implies evaluation bias. *Philosophical Studies*, *177*(12), 3661–3678. https://doi.org/10.1007/s11098-019-01387-w

Di Nardo, P. A., O'Brien, G. T., Barlow, D. H., Waddell, M. T., & Blanchard, E. B. (1983). Reliability of DSM-III Anxiety Disorder Categories Using a New Structured Interview. *Archives of General Psychiatry*, *40*(10), 1070–1074. https://doi.org/10.1001/archpsyc.1983.01790090032005

Dror, I. E., & Gallogly, D. P. (1999). Computational analyses in cognitive neuroscience: In defense of biological implausibility. *Psychonomic Bulletin & Review*, *6*(2), 173–182. https://doi.org/10.3758/BF03212325

Erdin, H. O. (2021). Appraisal of certain methodologies in cognitive science based on Lakatos's methodology of scientific research programmes. *Synthese*, *199*, 89–112. https://doi.org/10.1007/s11229-020-02612-4

Eronen, M. I., & Bringmann, L. F. (2021). The Theory Crisis in Psychology: How to Move Forward. *Perspectives on Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Ferguson, C. J., & Heene, M. (2012). A Vast Graveyard of Undead Theories: Publication Bias and Psychological Science's Aversion to the Null. *Perspectives on Psychological Science*, *7*(6), 555–561. https://doi.org/10.1177/1745691612459059

Fiedler, K. (1991). Heuristics and Biases in Theory Formation: On the Cognitive Processes of those Concerned with Cognitive Processes. *Theory & Psychology*, *1*(4), 407–430. https://doi.org/10.1177/0959354391014002

Fiedler, K. (2017). What Constitutes Strong Psychological Science? The (Neglected) Role of Diagnosticity and A Priori Theorizing: *Perspectives on Psychological Science*, *12*(1), 46–61. https://doi.org/10.1177/1745691616654458

Flis, I. (2019). Psychologists psychologizing scientific psychology: An epistemological reading of the replication crisis. *Theory & Psychology*, *29*(2), 158–181. https://doi.org/10.1177/0959354319835322

Frankenhuis, W. E., Panchanathan, K., & Smaldino, P. E. (2023). Strategic ambiguity in the social sciences. *Social Psychological Bulletin*. *18*, 1–25. https://doi.org/10.32872/spb.9923

Fried, E. I. (2020). Lack of Theory Building and Testing Impedes Progress in The Factor and Network Literature. *Psychological Inquiry*, *31*(4), 271–288. https://doi.org/10.1080/1047840X.2020.1853461

Frixione, M. (2001). Tractable competence. *Minds and Machines*, *11*, 379–397.

Gaj, N. (2016). *Unity and Fragmentation in Psychology: The Philosophical and Methodological Roots of the Discipline*. Taylor & Francis Ltd.

Giere, R. N., & Moffatt, B. (2003). Distributed Cognition: Where the Cognitive and the Social Merge. *Social Studies of Science*, *33*(2), 301–310. https://doi.org/10.1177/03063127030332017

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review*, *98*(2), 254–267. https://doi.org/10.1037/0033-295X.98.2.254

Gigerenzer, G. (1992). Discovery in Cognitive Psychology: New Tools Inspire New Theories. *Science in Context*, *5*(2), 329–350. https://doi.org/10.1017/S0269889700001216

Gigerenzer, G. (1998). Surrogates for Theories. *Theory & Psychology*, *8*(2), 195–204. https://doi.org/10.1177/0959354398082006

Gitelman, L. (Ed.). (2013). *"Raw data" is an oxymoron*. The MIT Press.

Goertzen, J. R. (2008). On the Possibility of Unification: The Reality and Nature of the Crisis in Psychology. *Theory & Psychology*, *18*(6), 829–852. https://doi.org/10.1177/0959354308097260

Gorelick, R. (2011). What is theory? *Ideas in Ecology and Evolution*, *4*, 1–10. https://doi.org/10.4033/iee.2011.4.1.c

Greenwald, A. G. (2012). There Is Nothing So Theoretical as a Good Method: *Perspectives on Psychological Science*, *7*(2), 99–108. https://doi.org/10.1177/1745691611434210

Greenwald, A. G., Pratkanis, A. R., Leippe, M. R., & Baumgardner, M. H. (1986). Under what conditions does theory obstruct research progress? *Psychological Review*, *93*(2), 216–229. https://doi.org/10.1037/0033-295X.93.2.216

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., … Zwienenberg, M. (2016). A Multilab Preregistered Replication of the Ego-Depletion Effect. *Perspectives on Psychological Science*, *11*(4), 546–573. https://doi.org/10.1177/1745691616652873

Hensel, W. M. (2020). Double trouble? The communication dimension of the reproducibility crisis in experimental psychology and neuroscience. *European Journal for Philosophy of Science*, *10*(3), 44. https://doi.org/10.1007/s13194-020-00317-6

Hensel, W. M., Miłkowski, M., & Nowakowski, P. (2022). Without more theory, psychology will be a headless rider. *Behavioral and Brain Sciences*, *45*, e20. https://doi.org/10.1017/S0140525X21000212

Hoyningen-Huene, P. (2013). *Systematicity: The Nature of Science*. Oxford University Press.

Hughes, B. M. (2018). *Psychology in crisis*. Palgrave.

Hutchins, E. (1995). *Cognition in the wild*. MIT Press.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Irvine, E. (2021). The Role of Replication Studies in Theory Building. *Perspectives on Psychological Science*, *16*(4), 844–853. https://doi.org/10.1177/1745691620970558

Isaac, A. M. C. (2019). Epistemic Loops and Measurement Realism. *Philosophy of Science*, *86*(5), 930–941. https://doi.org/10.1086/705476

Ivani, S. (2019). What we (should) talk about when we talk about fruitfulness. *European Journal for Philosophy of Science*, *9*(4), 1–18. https://doi.org/10.1007/s13194-018-0231-7

Kawa, S., & Giordano, J. (2012). A brief historicity of the Diagnostic and Statistical Manual of Mental Disorders: Issues and implications for the future of psychiatric canon and practice. *Philosophy, Ethics, and Humanities in Medicine: PEHM*, *7*, 2. https://doi.org/10.1186/1747-5341-7-2

Keas, M. N. (2018). Systematizing the theoretical virtues. *Synthese*, *195*(6), 2761–2793. https://doi.org/10.1007/s11229-017-1355-6

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., … Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*, 7–36. https://doi.org/10.1007/s40607-014-0009-9

Klein, S. B. (2014). What can recent replication failures tell us about the theoretical commitments of psychology? *Theory & Psychology*, *24*(3), 326–338. https://doi.org/10.1177/0959354314529616

Koyré, A. (1953). An Experiment in Measurement. *Proceedings of the American Philosophical Society*, *97*(2), 222–237.

Kuhn, T. S. (1977). *The essential tension: Selected studies in scientific tradition and change*. The University of Chicago Press.

Langley, P., Simon, H. A., Bradshaw, G. L., & Żytkow, J. M. (1987). *Scientific discovery: Computational explorations of the creative processes*. MIT Press.

Laudan, L. (1984). *Science and values: The aims of science and their role in scientific debate*. University of California Press.

Lean, O. M., Rivelli, L., & Pence, C. H. (2021). Digital Literature Analysis for Empirical Philosophy of Science. *The British Journal for the Philosophy of Science*. https://doi.org/10.1086/715049

Leonelli, S. (2016). *Data-centric biology: A philosophical study*. The University of Chicago Press.

Levenstein, D., Alvarez, V. A., Amarasingham, A., Azab, H., Chen, Z. S., Gerkin, R. C., … Redish, A. D. (2023). On the Role of Theory and Modeling in Neuroscience. *Journal of Neuroscience*, *43*(7), 1074–1088. https://doi.org/10.1523/JNEUROSCI.1179-22.2022

Litwin, P., & Miłkowski, M. (2020). Unification by Fiat: Arrested Development of Predictive Processing. *Cognitive Science*, *44*(7), e12867. https://doi.org/10.1111/cogs.12867

Longino, H. E. (1996). Cognitive and Non-Cognitive Values in Science: Rethinking the Dichotomy. In L. H. Nelson & J. Nelson (Eds.), *Feminism, Science, and the Philosophy of Science* (pp. 39–58). Springer Netherlands. https://doi.org/10.1007/978-94-009-1742-2_3

MacCorquodale, K. (1970). On Chomsky's review of Skinner's Verbal Behavior. *Journal of the Experimental Analysis of Behavior*, *13*(1), 83–99. https://doi.org/10.1901/jeab.1970.13-83

Manninen, T., Aćimović, J., Havela, R., Teppola, H., & Linne, M.-L. (2018). Challenges in Reproducibility, Replicability, and Comparability of Computational Models and Tools for Neuronal and Glial Networks, Cells, and Subcellular Structures. *Frontiers in Neuroinformatics*, *12*, 20. https://doi.org/10.3389/fninf.2018.00020

Marr, D. (1982). *Vision*. W. H. Freeman and Company.

Matthews, G. (2020). Against consensus: Embracing the disunity of personality theory. *Personality and Individual Differences*, *152*, 109535. https://doi.org/10.1016/j.paid.2019.109535

McMullin, E. (2010). The Virtues of a Good Theory. In M. Curd & M. Psillos (Eds.), *The Routledge Companion to Philosophy of Science* (pp. 561–571). Routledge. https://doi.org/10.4324/9780203744857.ch53

Meehl, P. E. (1967). Theory-Testing in Psychology and Physics: A Methodological Paradox. *Philosophy of Science*, *34*(2), 103–115. https://doi.org/10.1086/288135

Miłkowski, M. (2019). Fallible Heuristics and Evaluation of Research Traditions. The Case of Embodied Cognition. *Ruch Filozoficzny*, *75*(2), 223–236. https://doi.org/10.12775/RF.2019.031

Miłkowski, M. (2022). Cognitive Artifacts and Their Virtues in Scientific Practice. *Studies in Logic, Grammar and Rhetoric*, *67*(3), 219–246. https://doi.org/10.2478/slgr-2022-0012

Miłkowski, M., Hensel, W. M., & Hohol, M. (2018). Replicability or reproducibility? On the replication crisis in computational neuroscience and sharing only relevant detail.

Journal of Computational Neuroscience, *45*(3), 163–172. https://doi.org/10.1007/s10827-018-0702-z

Miłkowski, M., & Litwin, P. (2022). Testable or bust: Theoretical lessons for predictive processing. *Synthese*, *200*(6), 462. https://doi.org/10.1007/s11229-022-03891-9

Mischel, W. (2008). The Toothbrush Problem. *APS Observer*, *21*(11). https://www.psychologicalscience.org/observer/the-toothbrush-problem

Moretti, F. (2000). Conjectures on World Literature. *New Left Review*, *1*, 54–68.

Morgan, M. S., & Morrison, M. (1999). *Models As Mediators*. Cambridge University Press.

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*, 221–229. https://doi.org/10.1038/s41562-018-0522-1

Nersessian, N. J. (2008). *Creating scientific concepts*. MIT Press.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual information processing* (pp. 283–308). Academic Press.

Newell, A., & Simon, H. A. (1972). *Human Problem Solving*. Prentice-Hall.

Nickles, T. (2018). TTT: A Fast Heuristic to New Theories? In D. Danks & E. Ippoliti (Eds.), *Building Theories: Heuristics and Hypotheses in Sciences* (pp. 169–189). Springer. https://doi.org/10.1007/978-3-319-72787-5_9

Norman, D. A. (1991). Cognitive Artifacts. In J. M. Carroll (Ed.), *Designing Interaction: Psychology at the Human-Computer Interface* (pp. 17–38). Cambridge University Press.

Norton, J. D. (2021). *The Material Theory of Induction*. University of Calgary Press.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., … Vazire, S. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology*, *73*(1), 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://doi.org/10.1126/science.aac4716

Osbeck, L. M., & Nersessian, N. J. (2014). Situating distributed cognition. *Philosophical Psychology*, *27*(1), 82–97. https://doi.org/10.1080/09515089.2013.829384

Pence, C. H., & Ramsey, G. (2018). How to Do Digital Philosophy of Science. *Philosophy of Science*, *85*(5), 930–941. https://doi.org/10.1086/699697

Piper, A. (2020). *Can We Be Wrong? The Problem of Textual Evidence in a Time of Data*. Cambridge University Press.

Poldrack, R. A., Kittur, A., Kalar, D., Miller, E., Seppa, C., Gil, Y., … Bilder, R. M. (2011). The Cognitive Atlas: Toward a Knowledge Foundation for Cognitive Neuroscience. *Frontiers in Neuroinformatics*, *5*. https://doi.org/10.3389/fninf.2011.00017

Rescher, N. (1979). *Cognitive systematization: A systems-theoretic approach to a coherentist theory of knowledge*. Basil Blackwell.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358–358.

Rosenhan, D. L. (1973). On Being Sane in Insane Places. *Science*, *179*(4070), 250–258. https://doi.org/10.1126/science.179.4070.250

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2020). Why Hypothesis Testers Should Spend Less Time Testing Hypotheses. *Perspectives on Psychological Science*, *16*(4), 744–755. https://doi.org/10.1177/1745691620966795

Schindler, S. (2018). *Theoretical virtues in science: Uncovering reality through theory*. Cambridge University Press.

Schooler, J. W. (2014). Metascience could rescue the 'replication crisis.' *Nature*, *515*(7525), 9. https://doi.org/10.1038/515009a

Scull, A. (2023). Rosenhan revisited: Successful scientific fraud. *History of Psychiatry*, 0957154X221150878. https://doi.org/10.1177/0957154X221150878

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Shmueli, G., & Koppius, O. R. (2011). Predictive Analytics in Information Systems Research. *MIS Quarterly*, *35*(3), 553. https://doi.org/10.2307/23042796

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., … Nosek, B. A. (2018). Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Smaldino, P. E. (2017). Models Are Stupid, and We Need More of Them. In R. R. Vallacher, S. J. Read, & A. Nowak (Eds.), *Computational Social Psychology* (1st ed., pp. 311–331). Routledge. https://doi.org/10.4324/9781315173726-14

Staats, A. W. (1986). Unified Positivism: A Philosophy for Psychology and the Disunified Sciences. *Theoretical & Philosophical Psychology*, *6*(2), 77–90. https://doi.org/10.1037/h0091427

Sullivan, J. A. (2009). The multiplicity of experimental protocols: A challenge to reductionist and non-reductionist models of the unity of neuroscience. *Synthese*, *167*(3), 511–539. https://doi.org/10.1007/s11229-008-9389-4

Suppes, P. (1962). Models of Data. In E. Nagel, P. Suppes, & A. Tarski (Eds.), *Logic, Methodology, and Philosophy of Science: Proceedings of the 1960 International Congress* (pp. 252–261). Stanford University Press.

Szollosi, A., & Donkin, C. (2019). Neglected Sources of Flexibility in Psychological Theories: From Replicability to Good Explanations. *Computational Brain & Behavior*, *2*(3–4), 190–192. https://doi.org/10.1007/s42113-019-00045-y

Thagard, P. (1993). *Computational philosophy of science*. MIT Press.

Thagard, P., & Findlay, S. (2012). *The cognitive science of science: Explanation, discovery, and conceptual change*. MIT Press.

Trafimow, D., & Earp, B. D. (2016). Badly specified theories are not responsible for the replication crisis in social psychology: Comment on Klein. *Theory & Psychology*, *26*(4), 540–548. https://doi.org/10.1177/0959354316637136

Vadillo, M. A. (2019). Ego depletion may disappear by 2020. *Social Psychology*, *50*, 282–291. https://doi.org/10.1027/1864-9335/a000375

Vadillo, M. A., Gold, N., & Osman, M. (2016). The Bitter Truth About Sugar and Willpower: The Limited Evidential Value of the Glucose Model of Ego Depletion. *Psychological Science*, *27*(9), 1207–1214. https://doi.org/10.1177/0956797616654911

Van Rooij, I. (2008). The Tractable Cognition Thesis. *Cognitive Science*, *32*(6), 939–984. https://doi.org/10.1080/03640210801897856

Van Rooij, I., & Baggio, G. (2021). Theory Before the Test: How to Build High-Verisimilitude Explanatory Theories in Psychological Science. *Perspectives on Psychological Science*, *16*(4), 682–697. https://doi.org/10.1177/1745691620970604

Vohs, K. D., Schmeichel, B. J., Lohmann, S., Gronau, Q. F., Finley, A. J., Ainsworth, S. E., … Albarracín, D. (2021). A Multisite Preregistered Paradigmatic Test of the Ego-Depletion Effect. *Psychological Science*. (Sage CA: Los Angeles, CA). https://doi.org/10.1177/0956797621989733

Wilson, M. (1993). DSM-III and the transformation of American psychiatry: A history. *The American Journal of Psychiatry*, *150*(3), 399–410. https://doi.org/10.1176/ajp.150.3.399

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, e1. https://doi.org/10.1017/S0140525X20001685

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Young, G. (2016). *Unifying Causality and Psychology*. Springer International Publishing. https://doi.org/10.1007/978-3-319-24094-7

Zhang, J., & Norman, D. A. (1994). Representations in Distributed Cognitive Tasks. *Cognitive Science*, *18*(1), 87–122. https://doi.org/10.1207/s15516709cog1801_3

Zittoun, T., Gillespie, A., & Cornish, F. (2009). Fragmentation or Differentiation: Questioning the Crisis in Psychology. *Integrative Psychological and Behavioral Science*, *43*(2), 104–115. https://doi.org/10.1007/s12124-008-9083-6

Acknowledgements

The author wishes to thank Witold Hensel, Szymon Miłkoś, Przemysław R. Nowakowski for their comments to the draft of this paper, as well as to two reviewers of *The Review of Psychology*, Gustav Nilsonne and Sven Arend Ulpts, for their helpful referee reports.