Tribalism in Scientific Practice:

On the Failure to Erase Collective Misbeliefs in Science

Klaus Fiedler

Running head: Tribalsm in science

Author Note: The research documented in the present chapter was supported by research grant Fi 294/30-1 awarded by the Deutsche Forschungsgemeinschaft.

Abstract

Scientific evidence and rigor are commonly assumed to afford an appropriate remedy to the hazards of fake news and uncritical or even naïve misbeliefs. Yet, the present chapter shows that science is by no means immune to such anomalies. Rather, the scientific community, or which we all are a part, believes in a variety of collectively transmitted misbeliefs and false behavioral conventions, the invalidity of which is uncontested and well understood by a vast majority of scientists, who nevertheless continue to follow and normatively reify the misbeliefs. Prominent examples of such collective anomalies, which meet all defining features of tribalism in scientific practice, include ignorance of regressive shrinkage (e.g. in replication science), neglect of manipulation checks, the confident distinction of confirmatory and exploratory research, continued reliance on null-hypothesis significance testing and statistical power control, and the strict discrimination of experimental from correlational research. The concluding section is devoted to discussing possible explanation and remedies of these conspicuous anomalies.

Introduction

The term tribalism is charged with emotional and abnormal connotations and although it refers to a very common phenomenon, namely, collectively acquired beliefs, which are often plainly wrong but nevertheless distributed widely, due to conformity and compliance with naively adopted norms. The definition shared by authors of the present volume – tribalism as collectively acquired delusions – is not restricted to tribal lifestyles emerging in small hunter-gatherer groups as suggested in https://en.wikipedia.org/wiki/Tribalism. The present volume is rather concerned with collective beliefs and misbeliefs shared by connected communities of any size, including widely spread vocational communities representing scientific disciplines.

Plainly Wrong Myths in Science

Even members of intellectual circles hold a variety of conventions and habits, which are obviously wrong, sometimes to a blatant degree, but which are nevertheless adopted widely, as norms that govern study curricula, best-practice workshops, peer reviewing, and scientific evaluations. While the extant literature on fake news has mainly focused on a political instrument in the social media (Alcott & Gentzkow, 2017), serving self-confirmatory or outgroup-derogative purposes (Li & Su, 2020), the persistence of tribalism in science is independent of and often more general than suggested by these motivational accounts. Regardless of the origin of collective misbeliefs – be it self-presentation, social identity, conformity, or shallow thinking – their persistence seems to reflect a conspicuous metacognitive deficit in monitoring and controlling (correcting) judgments and inferences (Helson & Narens, 1994), the validity of which is either highly dubious or demonstrably wrong (Laughlin & Ellis, 1986). Such a naive deficit in meta-reasoning (termed “metacognitive myopia”; Fiedler, 2012; Fiedler, Prager & McCaughey, 2023) must be common to all persistent manifestations of all tribal misbeliefs, regardless of their origin. Given that most people, and scientists in particular, easily understand that certain beliefs are wrong, the most challenging question is why collective wisdom (i.e., the wisdom of crowds, Surowiecki, 2004) does not erase such misbeliefs. With reference to tribalism in science, we believe that collective misbeliefs are often manifestations of uncritical conformity or naïve compliance with ethical or methodological standards induced by epistemic authorities. These reflect the same weakness that the Jewish philosopher Hannah Ahrendt (1963) had in mind in her warning of conformity and her reminder of the obligation to be disobedient. Accordingly, the present compliance culture – with many people begging for politicians, physicians, experts, and administrators to tell them what they have to do – reflects the same weakness as the uncritical fellow-runners of dangerous movements in history. The kind of unwarranted myths in science that we shall tackle in this chapter could have never arisen and survived if a large sample of independent judges – especially intellectuals and scientists – would separately check for the validity and demonstrability (Laughlin & Ellis, 1986) of wrong beliefs. They could only arise because conformism prevents people from critical assessment and exploitation of the collective wisdom.

Collective delusions constitute a timely topic of research not only because people are so keen on compliance and so unlikely to engage in critical validity checks, which suggest themselves, but also because paternalism flourishes in these days. Ordinary people (consumers, voters, patients etc.) not only like it to be conformist and obedient; they are also expected to follow top-down instructions by those politicians, experts, and overlord decision makers who are supposed to know what is good for them. It is this paternalistic division of the population into origins and pawns (DeCharms, Karpenter & Kuperman, 1965), experts and laypeople, holding supposedly valid and invalid information, that justifies contrasts of valid and invalid evidence, news and fake news. The notion of "false and misleading information" presupposes that expert scientists at least understand what information is valid and true, rather than false and misleading.

Plainly Wrong Beliefs in Replication Science

In any case, whether or not our intuition is correct that persistent misbeliefs in science reflect paternalism and the compliance culture, the phenomenon seems obvious, as illustrated by few initial examples from the last decade of intensive replication science. There is wide consensus from several frequently cited replication programs, suggesting that the majority of replication attempts are not met with success. Apart from the significance criterion, replication projects generally find smaller effect sizes in replications than in the initial version of the same investigations, as shown in the Open Science Collaboration's (2015) systematic attempt to replicate a sample of 100 experiments published in major journal. Virtually every young student at undergraduate level, every journalist, and every young scientist involved in the peer reviewing of international journals "knows" that most replication effect sizes are regularly weaker than "original" effect sizes. Yet, hardly anybody "knows" that this shrinkage of effect sizes is trivial and does not reflect any novel insight, unless the observed reduction in effect size is shown to be stronger than the expected regressive shrinkage. It can be expected on normative grounds.

Regressive shrinkage. An eye-opening simulation study by Stanley and Spence (2014) shows that measurement error alone guarantees that 100,000 replications of the same *N* participants from the same population with a known effect size of *r* yields a widely dispersed distribution of effect sizes. Given a reliability of $\rho = .70$, a sample size of $N = 40$, and a true population effect size of $r = .30$, replicated effect sizes vary markedly between .00 and .50. It is a completely normal, trivial, property of the empirical world that nuisance alone (due to measurement error) leads to highly variable results, the precise dispersion of which depends on *N*, ρ, and *r* in a predictable way.

Moreover, it is completely normal that variation between multiple replications are subject to regressive shrinkage (Campbell & Kenny, 1999; Fiedler & Krueger, 2012). Replication effects tend to be lower than high original effect sizes ($r > 0$) but higher than low original effect sizes ($r < 0$), and this regression effect is more pronounced when effect sizes

($|r|$) are extreme, reliabilities ρ are restricted, and sample sizes *N* are low. Yet as long as the correlation between initial and replicated effect sizes is less than perfect, regressive shrinkage is inevitable. To quote Campbell and Kenny (1999, p. ix), it is "as inevitable as death and taxes".

Thus, the null hypothesis of virtually all replication science – assuming equal effect sizes for replicated and original studies – is logically unwarranted and the "finding" that replication effects are weaker than the "original" effect sizes constitutes an empty truism. Selling it as a "finding" is misleading and reflective of shallow, unscientific reasoning. Because the expected (original) effect size is inevitably subject to regression, it would be necessary to estimate the expected amount of "normal" regression. Every empirical scientist with elementary training in statistics ought to know in advance that replicated findings cannot be as extreme as original findings. The phenomenon of inverse regression (Baltes, Nesselroade, Schaie & Labouvie, 1972) means that when plotting original effect sizes as a function of replicated effect sizes, the result is regressive too. Thus, large replication effects tend to be smaller in the "original", whereas the original size of small replication effects tends to be larger, as shown by Fiedler & Prager (2018) with reference to the OSC results. This "surprising" finding is trivial. It is analogous to the truism that an imperfect correlation implies a slope <1 both for the regression of Y on X and for the regression of X on Y.

Reminding empirical scientists of the truism of regression is typically met with anger and responses like "you need not explain me what I learned in my first study year". And yet, insensitivity to regressive shrinkage has hindered scientific reasoning for over hundred years (Baltes et al., 1972; Furby, 1973; Galton, 1886; Rulon, 1941). The notion of regressive shrinkage – although it is as certain as death and paying taxes – is persistently ignored in replication science. To the best of our knowledge, replication researchers have never made a systematic attempt to rule out regression in estimates of the expected ("true") effect size when plotting observed replication effects as a function of observed initial effect. It is as if the basic notion of regression does not belong to the tribal knowledge frame of replication science. To be sure, the problem is not lack of analytic skills to calculate regression effects. It rather reflects a deeply wired unwillingness or lack of motivation to beware of the "regression trap", which is at the heart of a whole class of serious scientific misbeliefs. The majority of scientists overlook the regression trap although they easily understand that regressive shrinkage is ubiquitous.

What empirical findings can be expected to be replicable? In the area of replication science, we encounter several other conspicuous examples of silently adopted counter-facts, established through frequent usage but devoid of any evidence. Thus, it is commonly taken for granted that all published findings are equally relevant candidates for replication, as evident in the fact that replication studies are selected randomly from the entirety of all studies published in certain journals (Camerer et al. 2018; OCE, 2015). The mere possibility that not all hypothesis tests are meant to deliver replicable findings (see Fiedler, 2017) is hardly ever considered, although this possibility is obvious. For instance, why should Pfungst's (1965) famous research on Clever Hans (i.e., an extraordinary horse that could apparently calculate) be replicable, or Schaller and Park's (2011) ingenious work on the moderating impact of disgust stimuli on immune reactions, or findings for which crucial auxiliary assumptions (e.g., classical conditioning after blocking) are not met (see Fiedler & Trafimow, 2023)?

The logical function of manipulation checks. Still another compelling example is the ignorance of manipulation checks (Fiedler, McCaughey & Prager 2021). If you want to test the hypothesis that a certain substance is lethal, you may count whether the number of dead creatures increases after the consumption of the poisonous substance. Yet, if the substance was not eaten, the hypothesis cannot be tested. Nobody would pretend this to falsify the hypothesis, when the substance was not eaten.

More generally, if an experiment is designed to test the hypothesis that a manipulation in an independent variable, ΔX, causes a change in a dependent variable, ΔY, a failure to demonstrate or to replicate a finding need not be due to invalidity of the causal relation $\Delta X \rightarrow \Delta Y$. It can as well reflect the failure to induce the critical ΔX shift in the independent variable. In the absence of a check on ΔX, a count of ΔY cannot tell us anything about the validity of the hypothesis. In other words, a manipulation check is logically essential and, because replication research is meant as a critical quality check, such a manipulation check is even more essential for replications than for initial research. Without manipulation check, a failure to replicate remains totally ambiguous; it could simply reflect too weak and inappropriate an attempt to manipulate the independent variable, as a premise to the hypothesis test. Yet, manipulation checks are not a necessary condition of so far published replication research, though their neglect undermines a logical rationale of replication (Fiedler & Ermark, in press).

Significance testing: Survival of a Flawed Method

The common habit of null hypothesis significance testing (NHST) characterizes the methodology of virtually all research in behavioral science. Both experimental and correlational research are usually framed as significance tests comparing a focal hypothesis (postulating a critical finding) with a null hypothesis (admitting no finding). Fulfilling the pre-conditions of a significance test as the aim of virtually all so-called good practices of science and virtually all debates about ethical versus unethical conduct (John, Loewenstein & Prelec, 2012; Simmons, Nelson & Simonsohn, 2011). Unfortunately, the holy cow of significance-testing is unwarranted, according to the long-known insight that NHST is elusive, reflecting the survival or a flawed method (Cohen, 1994; Hunter, 1997; Krueger, 2001).

Confusion of the conditional inference direction. Logically, NHST presupposes the conditional probability $p(D \mid H_0)$ of obtaining a certain data pattern D given the null hypothesis H_0 that the true effect is exactly zero. However, the aim of inferential statistics is to infer the inverse conditional $p(H_0 \mid D)$ of a hypothesis (H_0 or H_1) given an observed data set D.

Bayesian calculus shows that the conditional probability $p(H_0 | D)$ of a hypothesis H_0 given a data pattern D is the reverse probability $p(D | H_0)$ of data given the (null) hypothesis multiplied with the ratio of base-rates, $p(H_0)/p(D)$. Unfortunately, however, these base-rates are unknown in reality, making it impossible to infer the conditional probability of a hypothesis of either (H_0 or H_1=1–H_0) from the data.

Indeed, the conditions of NHST inferences are practically never met (Wason, 1965). Even formally educated researchers virtually never consider the underlying assumptions. They understand that NHST focuses on the conditional probability $p(D | H_0)$ to obtain a data set (at least) as strong as D, given the null hypothesis $p(H_0)$ that the true difference is exactly zero. They also understand that this is virtually never the case. How likely it is in reality that a true effect size is exactly zero, given that all real variables are somehow correlated? To give meaning to such an absurd situation, the commit a serious reverse inference mistake, treating highly significant results [less unlikely than expected at a small error probability $\alpha = p(D | H_0)$] as if the inverse probability $p(H_0 | D)$ is also small and, by complement, the likelihood $p(H_1 | D)$ of the focal hypothesis given the data is “significant”. This reverse inference is of course wrong.

As a consequence of the obvious fact that H_0 refers to an extremely unlikely state of nature, it is also obvious that the greatest part of false positives (i.e., data that overestimate the true effect) result from situations in which H_0 is not met. As a consequence, the actual false-positive rate must be much higher than $\alpha = p(D | H_0) = .05$ (see Hunter, 1997). An appropriate Bayesian estimate of the false positive rate must consider many other conditions except the highly unlikely H_0. Many researchers and formally educated statisticians understand that NHST is a phantom, the domain of which approximates zero. They nevertheless continue to advocate and teach NHST and to base virtually all their published work and most of their journal reviews on normative NHST rules. Rather than accepting that NHST is a flawed method, they resort to an unwarranted reverse inference, interpreting a numerically low α as if the $p(H_0 | D)$ is also low, whereas the likelihood of the focal hypothesis $p(H_1 | D)$ is high. This sort of an elusive reverse inference is almost obligatory in the tribal NHST community of behavioral science.

Illusion of power control. The elusive status of the NHST industry is even more evident when it comes to estimating the false-negative error rate β and its complement, the statistical power $1–\beta$, that is, the hit rate of obtaining significant evidence for H_1 when H_1 is actually true. Almost all empirical journals call for a test, or at least for a discussion of statistical power, when authors want a submitting paper to pass the peer reviewing process. To comply with this obligatory norm, authors typically estimate the statistical power of their hypothesis test using a software tool, such as GPower (Faul, Erdfelder, Lang, & Buchner, 2007). Based on an effect-size estimate of the major hypothesis, the software calculates the number of participants (per cell) required to obtain a significant finding at a power of, say, 90% ($1–\beta = .90$). In other words, the aim of the collectively enforced power prelude is to determine the participant sample size required to guarantee a significant result (i.e., a “hit”) in case that H_1 is actually true (i.e., that the population effect size is different from zero, so that H_0 is false). Indeed, the authors of virtually all articles, even in leading journals, believe and pretend that the power promised by such a software tool actually provides a lower limit for the study at hand. They assume the likelihood of true effect to be borne out in the study at hand is at least $1–\beta = .90$.

To understand that this common power inference qualifies as a collective delusion self-deception, consider the following thought experiment. You have already “secured” the power of a new experiment, using GPower or some other software tool, relying on the median effect size obtained in a meta-analysis of similar past experiments. Now you are about to conduct the experiment tomorrow. However, before the experiment really starts, you (or maybe your student experimenters) change several features that are completely irrelevant for GPower. For instance, the title of the announced experiment may change, the instructions, the presentation speed, the study language, the experiment is conducted at night, in serious fatigue, or maybe under alcohol, participant are underpaid (in MTurk), or the experiment is the last in a demanding session of six other experiments, etc. Although all these changing boundary conditions will not leave the effective power of the new experiment unaffected, as everybody can understand, without any formal training, virtually all researchers ignore this insight and obediently participate in the elusive "power discussion" game (Fiedler, 2020).

A disarming lesson of this thought experiment is that one should never assume that a statistical power estimate holds for the specific experiment one is currently conducting. As Judd and Kenny (in press) have vividly shown, statistical power estimates and replicability forecasts exclusively consider the impact of sampling error due to random factors within the design. They are insensitive to constant boundary conditions (such as study language, presentation speed, fatigue) that do not vary within experiments across the levels of a random factor, but only between experiments or task settings. Because of these uncontrolled boundary conditions, the effective power and replicability are regularly lower than estimated statistically. It is hard if not impossible to quantify the resulting estimation error. The collective belief in, and the enforcement or control of β constitutes a tribally established delusion par excellence.

Pitfalls of Research Designs

Let us now discuss uncritically accepted assumptions outside the domain of statistics, related to collective delusions about research designs. These issues also highlight that tribalism is not peculiar to deficits in numeracy or formal reasoning. Its domain is much broader, covering non-statistical reasoning as well. A very prominent conceptual distinction to start with refers to the contrast of confirmatory and exploratory research, which is supposed to be a major index for the value of research.

The basic idea is that unlike confirmatory research concerned with isolated tests of a-priori hypotheses, exploratory research merely allows for incidental results observed in a-posteriori analyses of unpredicted findings. Related to this distinction are the distinctions between inferential and merely descriptive findings and the distinction between experimental and merely correlational research. Common to all these overlapping distinctions is the assumption that research varies dramatically in terms of theoretical predictability or constraints imposed on the results. The primary assumption states that confirmatory research is of superior quality, and much more informative, than merely exploratory research.

Confirmatory versus exploratory research. For a number of reasons, this assumption turns out to be untenable and simply wrong in many concrete cases. First, an investigation does not improve from exploratory to confirmatory just because it is preregistered. If a hypothesis test does not strictly derive logically or deductively under theoretical constraints, the bureaucratic act of preregistration does not make it confirmatory. Imagine a typical yin-and-yon lottery experiment with two possible outcomes. A treatment or intervention may let outcomes increase or decrease, make people happier or sadder, render an attitude stronger or weaker (Fiedler, 2017). The information value of such a dichotomous outcome is not higher than the prediction of either heads or tails in coin tossing. Conversely, if an investigation's results can be derived logically from a well-established theory, the a-priori nature of the results is definitely not contingent on preregistration.

Secondly, the terms "exploratory" and "confirmatory" are easily misunderstood; "exploratory" does not mean that no a-priori prediction exists, and "confirmatory" does not imply a confirmatory result. Because the confirmation of valid theoretical predictions (e.g., of a classical-conditioning effect) are contingent on auxiliary assumptions (e.g., that no blocking effects work against conditioning; Fiedler & Trafimow, 2023), a clearly confirmatory approach (based on a fully valid and correct mechanism of conditioning and blocking) may predict disconfirmation when blocking is at work. At the same time, perfect theoretical understanding of the relative strength of conditioning and blocking may render the outcome of a conditioning study undecided and exploratory.

A third reason lies in the important fact that no experimental intervention constitutes a plain manipulation of only one independent variable. For example, manipulating stimulus presentation time sounds like a clear-cut manipulation of exactly one physically defined variable: presentation time, which can be measured objectively in milliseconds. However, psychologically, the manipulation of presentation time can affect a variety of psychologically different but confounded causes, suggesting fundamentally different theoretical accounts of the same experimentally induced effects. In addition to affecting stimulus exposure time or speed, the manipulation of presentation time may induce time pressure, emotional stress, fatigue, or speed-dependent processing strategies, to list but a few.

Likewise, a persuasive communication supposed to induce unequal proportions of pro and contra arguments may also induce unwanted demand effects, defensive or hostile emotional reactions, influence processing depth, re-activate autobiographical memory traces, etc. As a rule, there is hardly any experimental treatment or intervention that constitutes a pure manipulation of an experiment's focal independent variable. Virtually every intervention causes a joint manipulation of a variety of correlated attributes, thus blurring the distinction between a randomized experiment and merely correlational research.

Experimental versus correlational research. More generally, the apparently tight distinction between experimental and merely correlational design is, most of the time, elusive. It qualifies as another example of misbelief in science. For the aforementioned reasons, an experiment that relies heavily on the isolated manipulation of one intended factor, does not guarantee that other, unintended factors are all cancelled out through randomization. Rather, the manipulated factor is inevitably correlated with a whole bunch of confounded influences.

To provide one more example, if an experimental condition and a control condition only differ in the manipulated construal level (i.e., abstractness) of the instruction text, construal-level theory (Trope & Liberman, 2003, 2010) tells us that the two experimental conditions vary simultaneously in many correlated ways. A high construal-level condition will induce higher spatial, temporal, social and hypothetical distance from the stimuli than a concrete condition, a more multidimensional representation, relatively higher weight given to outcome value than to probabilities, and many other factors correlated with construal level. Any decision about which one of these confounded causal factors accounts for the variance observed in the dependent variable (e.g., strength of fundamental attribution bias, which typically increases from low to high construal level; Nussbaum, Trope & Liberman, 2003) must rely on post-hoc speculations on the plausibility of multiple correlated attributes. Because there is no such thing as a truly isolated manipulation of one unique independent variable, all research is to some degree correlational.

The same lesson follows from a moment of reflection on attrition rates. Willingness to participate in, and to complete, an experiment on a certain topic is not at all random but contingent on lots of hedonic and motivational factors. The attrition rate of MTurk participants is often over 40%, and it is by no means constant in the experimental and the control condition (Zhou & Fishbach, 2016). Thus, contrary to the basic assumptions of a randomized design, participants who complete a demanding experimental condition are often more highly motivated or more resilient than participants in a less demanding control condition. While MTurk does not spontaneously inform researchers about the effective attrition rates, the rules of the tribal game rarely encourage researchers to ask for the attrition rates. They rather stick to the elusive assumption of participants as a random factor.

Fixed versus random effects. The term "random factor" is actually quite fashionable; it is propagated as a precondition to render findings generalizable not only across participants but also across tasks, trials, stimulus faces treated as random factors (Judd, Westfall & Kenny, 2012). Many well-motivated scholars recommend mixed models to improve the statistical data analysis and to approximate the ideal of a representative design (Brunswik, 1955; Dhami, Hertwig & Hoffrage, 2004). Including more random factors in a design should indeed increase the generalizability of results. Yet, a true random factor must first of all be a genuine random factor in the research design. It is not enough to treat a non-random factor as random in statistical analyses. Recall that for a finding to be generalizable over a random factor, the factor levels must represent a random sample of the factor's distribution in the population. Thus, if the design of an overconfidence experiment includes the same 50% proportion of hard and easy judgment tasks, although the natural distribution of hard and easy tasks is highly skewed (with hard tasks much more infrequent than easy tasks), treating tasks as random factor will hardly result in generalizable results. The distorted skew of a "random factor" may greatly obscure the typical "hard-easy effect" (Juslin, 1994; Moore & Healy, 2008), that is, the asymmetry of relatively more overconfidence on hard problems but more under-confidence on easy problems, if the design strongly over-represents hard problems. Likewise, the orthogonal variation of two random factors, participant knowledge and task difficulty, may obscure their interaction, because in reality participant knowledge (expertise) and task difficulty may not be orthogonal at all. For various reasons, then, a design factor does not become random and does not allow for strong generalization simply because it is treated as a random factor in a mixed statistical analysis. It is rather essential that a random factor be treated as such in the study design.

Neglecting weaker alternatives in a hierarchy of accounts. Although most researchers are familiar with the lessons gained from Peter Wason (1960), they often fail to act accordingly. They do not seem to recognize that seemingly compelling evidence for the operation of a manipulated causal factor may as well reflect a much weaker causal interpretation, referring to a more abstract level in a hierarchy of nested causes. For a prominent example, inducing mortality salience (by exposing participants to a funeral or letting them write a brief essay about death) has been shown to induce a conservative shift in political and social preferences (Weise et al., 2008). The more frequently this finding was replicated or cross-validated, the stronger grew the researchers' conviction that mortality salience must be the causal origin of the conservative shift. However, in a hierarchy of causes, mortality fear is just a special case in a hierarchy of many other versions of incompleteness.

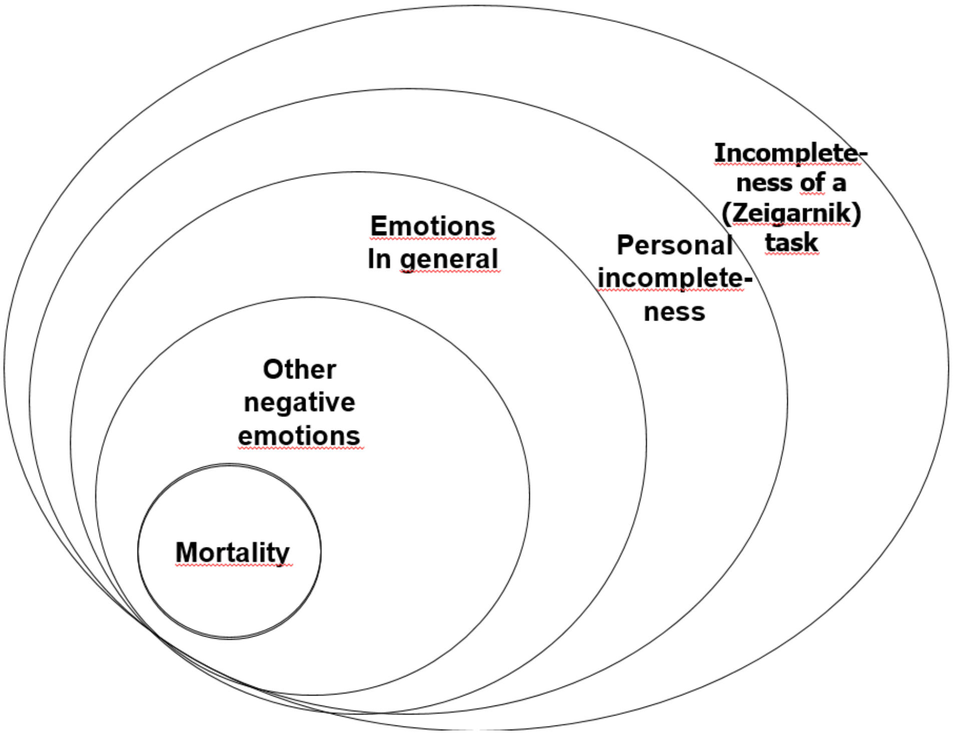

Figure 1: Hierarchy of increasingly weaker causal explanations of mortality salience effects, analogous to Wason (1960)

As evident from Figure 1, mortality is a special case of negative (fear-inducing) emotions, which are special cases of emotions (including positive joy after the birth of one's child), which may also induce a similar shift towards conservative values. Even more general is the feeling of incompleteness (as induced by the reminder that participants are in their beginning semesters), or even in a Zeigarnik (1938) effect induced by interrupting participants on an irrelevant task (fully unrelated to fear or even mortality), which may be sufficient to induce a mind-state related to incompleteness. Long-known research by Wicklund and Braun

(1990) suggests that incompleteness of such a much weaker kind may have similar effects as exposure to mortality, thus providing an alternative account of findings that have been traditionally attributed to the specific causal influence of mortality. After all, political media reports render the mortality account plausible, such as the conservative candidate George W. Bush winning the 2003 election against John Carey after the mortality-prone 9-11-catastrophy. More generally, it is often possible to conceive of a hierarchy of general versus concrete causes offering a variety of alternative accounts of the same empirical finding.

Mediation analysis. Finally, a common source of a shared elusive norm in current science is mediation analysis (Baron & Kenny, 1986; Judd & Kenny, 1981) – a method that "... is now almost mandatory for new social-psychology manuscripts" (Bullock, Green & Ha, 2010, p. 550) to be published in leading journals. The grand image of statistical mediation analysis reflects the widely shared conviction that it offers a statistical method to diagnose the causal mechanism underlying an influence X→Y of an independent variable C on a dependent variable Y. A statistical mediation test of the model X→M→Y is considered cogent evidence that variable M mediates the impact of X on Y.

For instance, if a recipient's attitude changes in the direction of a persuasive message, a mediation model may assume that the mediator M of an attitude shift Y induced by the persuasive message X may be the recipient's relative number of pro minus anti responses to the message. If exposure to the message (X) correlates positively with the tendency to produce pro (vs. anti) responses (M), and M correlates in turn with a corresponding attitude shift (Y), the regression of Y on X disappears or decreases when both X and M are jointly included in a regression model (i.e., $\beta_{XY.M} < \beta_{XY}$).

The common belief that this sort of a correlation pattern provides cogent evidence for a causal mediation model (X→M→Y) constitutes a widely shared tribal misbelief (Fiedler, Harris & Schott, 2018). To be sure, researchers "know" that correlational findings cannot inform causal inferences. They know that many other mediator variables exist that could also reduce the regression of Y on X (e.g., a demand effect induced by the persuasive message). Many understand that similar causal models (e.g., the common-cause model X,M→Y or the inverse mediation model X→Y→M) show a very similar covariance structure as the mediation model X→M→Y.

Computer simulations by Fiedler, Schott and Meiser (2011) have shown that when all variables are equally correlated in the population, any assignment of three randomly selected variables to the role of X, M, and Y will produce a positive meditation test (contingent only on sufficient sample size). The origin of the misbelief is that statistical mediation tests focus on a single causal model (e.g., mediation) and ignore the existence of (at least) 12 alternative causal models. Moreover, users of mediation analysis strongly underestimate the number of alternative mediation variables.

The elusive status of the mediation hype is evident in the fact that virtually 100% of all researchers using mediation analysis erroneously assume that a significant mediation test implies that they have diagnosed the causal mediator underlying their data (Fiedler et al., 2018).

Conclusions

To summarize, the present chapter has shown, as intended, that tribalism cannot be attributed to immature collective misbeliefs shared by unintelligent and uneducated people, who lack the meta-cognitive means of critical assessment. Rather, the collective adoption and maintenance of elusive beliefs can also be found in the scientific community – that is, in an intellectual elite of highly educated and enlightened minds that are professionally accustomed to critical assessment and evidence-based validation. In spite of strong arguments and uncontested demonstrations that many collectively shared scientific beliefs are unwarranted, and even when those counter-arguments are easy to understand and widely accepted, scientists continue to adhere to those misbeliefs, and they are reluctant to discard what is collectively embraced as normative standards. Thus, replication scientists continue to ignore regressive shrinkage, scientists neglect manipulation checks, overstate the distinctions of experimental and confirmatory versus correlational and exploratory science, naively trust in mixed designs and mediation results, and they are often blind for the fact that much weaker hypotheses can explain the same findings as the preferred strong hypotheses.

Two pertinent questions that suggest themselves are: First, why do smart scientists fall prey to elusive misbeliefs, the invalidity of which they can easily understand? And second, what remedies will most likely help critically-minded scientists to overcome this curious susceptibility to tribal beliefs?

Regarding the first question, we have already stated that meta-cognitive myopia (Fiedler, 2012; Fiedler et al., 2023) constitutes a necessary condition for the persistence of tribalism in science. Misbeliefs could hardly survive if they were subject to critical validity checks by smart scientists whose intelligence allows them to diagnose belief validity. It seems similarly clear that conformity must be at work. If only a minority of scientists really understand the logical principles that invalidate elusive beliefs, the wisdom of crowds should before too long cancel out wrong beliefs (Laughlin & Ellis, 1986), unless epistemic authorities prescribe elusive norms in a genuinely conformist process. Without such conformist compliance, it would be hardly possible that many individual scientists' independent (i.e., non-conformist) validity checks should reinforce existing myths and disbeliefs. Thus, even when many scientists do not fully understand the regression trap or fail to realize the impact of complex experimental manipulations and the importance of manipulation check, one would still expect a smart minority to dominate the long-term outcome.

There can be no doubt that tribalism profits from intuitive uncertainty. Thus, one cannot expect intuition and superficial everyday thoughts to cope with such non-intuitive principles as regressive shrinkage, critical manipulation checks, or random design factors. Rather, we believe that metacognitive quality checks (involving monitoring and control) and social validation processes that exploit the wisdom of crowd, are necessary to cope with these anomalies. As a consequence, metacognitive and interpersonal quality checks should also be the focus of any promising remedy intended to overcome at least the most obvious manifestations of tribalism in science.

We believe that the current compliance culture in science, which obliges young scientists to imitate established habits and prevents them from courageous critique and self-determined reasoning, facilitates both conformity and metacognitive myopia. Conversely, we believe that a reasonable way to overcome these conformist limitations and to get rid of collective elusions is to encourage pluralism and independent thinking in science and respect for those dissenters who dare to argue against the mainstream. Realistically, there can be no guarantee that tribalism can be mastered but the only reasonable way at least to ameliorate the problem is to work against the naïve mindset and the collective nature of tribalism, that is, to overcome metacognitive myopia and conformity in science.

References

Ahrendt, H. (1963). *Eichmann in Jerusalem: A Report on the Banality of Evil*, London, Faber & Faber.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211-236.

Baltes, P. B., Nesselroade, J. R., Schaie, K. W., & Labouvie, E. W. (1972). On the dilemma of regression effects in examining ability-level-related differentials in ontogenetic patterns of intelligence. *Developmental Psychology*, *6*(1), 78–84.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*(3), 193–217.

Bullock, J., Green, D., & Ha, S. (2010). Yes, but what's the mechanism? (don't expect an easy answer). *Journal of Personality and Social Psychology, 98*, 550–558.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... & Altmejd, A. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637-644

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts*. Guilford Press.

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist, 49*, 997-1003.

DeCharms, R., Carpenter, V., & Kuperman, A. (1965). The" origin-pawn" variable in person perception. *Sociometry*, 241-258.

Dhami, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin*, *130*(6), 959–988.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101,* 519–527.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). GPower 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Fiedler, K. (2011). Voodoo correlations are everywhere—Not only in neuroscience. *Perspectives on Psychological Science*, *6*(2), 163-171.

Fiedler, K. (2012). Meta-cognitive myopia and the dilemmas of inductive-statistical inference. In *Psychology of Learning and Motivation* (Vol. *57*, pp. 1-55). Academic Press.

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.

Fiedler, K. (2020). Elusive alpha and beta control in a multicausal world. *Basic and Applied Social Psychology. 42*(2), 79-87.

Fiedler, K., & Ermark, F. (2023). Replication in social and personality psychology. In H. T. Reis, T. West, and C.M. Judd (Eds.), *Handbook of Research Methods in Social and Personality Psychology*. Cambridge: Cambridge University Press (3rd edition)

Fiedler, K., Harris, C., & Schott, M. (2018). Unwarranted inferences from statistical mediation tests–An analysis of articles published in 2015. *Journal of Experimental Social Psychology*, *75*, 95-102.

Fiedler, K., & Krueger, J. I. (2012). More than an artifact: Regression as a theoretical construct. In J. I. Krueger (Ed.), *Social judgment and decision making.* (pp. 171–189). Psychology Press.

Fiedler, K., McCaughey, L., & Prager, J. (2021). Quo vadis, methodology? The key role of manipulation checks for validity control and quality of science. *Perspectives on Psychological Science*, *16*(4), 816–826.

Fiedler, K., & Prager, J. (2018). The regression trap and other pitfalls of replication science—Illustrated by the report of the Open Science Collaboration. *Basic and Applied Social Psychology*, *40*(3), 115–124.

Fiedler, K., Prager, J. & McCaughey, L. (2023). Metacognitive myopia: A major obstacle on the way to rationality. *Current Directions of Psychological Science*. https://doi.org/10.1177/09637214221126906.

Fiedler, K., Schott, M., & Meiser, T. (2011). What mediation analysis can (not) do. *Journal of Experimental Social Psychology*, *47*(6), 1231-1236.

Fiedler, K., & Trafimow, D. (2023). *Using and refining the TASI taxonomy to delineate predictably replicable findings.* Manuscript submitted for publication.

Furby, L. (1973). Interpreting regression toward the mean in developmental research. *Developmental Psychology, 8*, 172-179.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland, 15,* 264–263.

Hartwig, M., & Bond, C. F., Jr. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin*, *137*(4), 643–659.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, *8*(1), 3-7.

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.

Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review, 5*, 602–619.

Judd, C.M., & Kenny, D.A. (in press). Random Factors and Research Generalization. In H. T. Reis, T. West, and C.M. Judd (Eds.), *Handbook of Research Methods in Social and Personality Psychology*. Cambridge: Cambridge University Press (3rd edition)

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.

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes*, *57*(2), 226–246.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist*, *56*(1), 16-26.

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology*, *22*(3), 177-189.

Li, J., & Su, M. H. (2020). Real talk about fake news: Identity language and disconnected networks of the US public's "fake news" discourse on Twitter. *Social Media+ Society*, *6*(2), 2056305120916841.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517.

Nussbaum, S., Trope, Y., & Liberman, N. (2003). Creeping dispositionism: the temporal dynamics of behavior prediction. *Journal of Personality and Social Psychology, 84*(3), 485.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Pfungst, O. (1965). *Clever Hans: The horse of Mr Von Osten*. Holt, Rinehart & Winston, Inc.

Popper, K. (1959). *The logic of scientific discovery*. Oxford England: Basic Books.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*, 358–367.

Rulon, P. J. (1941). Problems of regression. *Harvard Educational Review, 11*, 213-223.

Schaller, M., & Park, J. H. (2011). The behavioral immune system (and why it matters). *Current Directions in Psychological Science*, *20*(2), 99–103.

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.

Trope, Y., & Liberman, N. (2003). Temporal construal. *Psychological Review*, *110*(3), 403–421.

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review*, *117*(2), 440–463.

Weise, D. R., Pyszczynski, T., Cox, C. R., Arndt, J., Greenberg, J., Solomon, S., & Kosloff, S. (2008). Interpersonal politics: The role of terror management and attachment processes in shaping political preferences. *Psychological Science*, *19*(5), 448–455.

Wicklund, R. A., & Braun, O. L. (1990). Creating consistency among pairs of traits: A bridge from social psychology to trait psychology. *Journal of Experimental Social Psychology*, *26*(6), 545–558.

Zeigarnik, B. (1938). On finished and unfinished tasks. In W. D. Ellis (Ed.), *A source book of Gestalt psychology.* (pp. 300–314). Kegan Paul, Trench, Trubner & Company.

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology, 111*(4), 493-504.