Ethical Norms and Moral Values Among Scientists:

Applying Conceptions of Morality to Scientific Rules and Practices

Klaus Fiedler

University of Heidelberg

To appear in:

Joseph P. Forgas, Paul van Lange & Lee Jussim (Eds.)

Social Psychology and Morality

New York: Psychology Press

Abstract

The purpose of the present paper is to apply theoretical approaches to the social psychology of morality to the moral evaluation of scientists' own behavior. It will be seen that the current discourse on questionable research practices, usability of science, and fraud, mainly fueled by whistle-blowers who are themselves members of the scientific community, does not live up to higher levels of moral judgment according to Piaget and Kohlberg. This discourse can also not be explained within the modern intuitionist approach to morality. It can however be understood in terms of the current Zeitgeist of compliance with unreflected norms, related to the social psychology of conformity and obedience. Transgressions of arbitrary norms (e.g., of sober significance testing) are given more weight than violations of more fundamental norms (e.g., theoretical reasoning; cost-benefit assessment). The age-old principle of reciprocity is largely neglected when powerful players in the game (editors', reviewers', whistle-blowers' motives or conflicts of interest) are not treated according to the same rules as weaker players (authors, students, applicants). A recent critique by Fiedler and Schwarz (2015) of John et al.'s (2012) evidence on questionable research practices is used for illustration. The final discussion is concerned with practical and theoretical implications of the moral values among scientists.

Introduction

According to the Stanford Encyclopedia of Philosophy[1], the term morality can be defined either descriptively "to refer to some codes of conduct put forward by a society" or normatively "to refer to a code of conduct that, given specified conditions, would be put forward by all rational persons." Traditionally, the seminal research by Piaget (1965), Kohlberg (1963), and Kelley (1971) on the psychology of morality was apparently driven by the latter, normative perspective, presupposing that morality constitutes an integral aspect of rationality. Jean Piaget's (1932) monograph conveys the message that the development of moral judgment in the child is closely linked to the maturation of intelligence. He was obviously influenced by Kant's (1788 / 1999) belief in the a-priori existence of both cognitive and moral categories, the ontogenetic acquisition of which Piaget believed to follow a regular sequence of developmental stages. This assumption can also be encountered in Lawrence Kohlberg's (1963) writings. From a slightly different, attributional perspective, Harold Kelley (1971) proposed: "*The moral evaluation process is, in part, based on the processes of reality evaluation and achievement evaluation.* By this, I mean that judgments of right and wrong, good or bad (moral evaluations), derive their properties in part from the same processes as are involved in judgments of correct or incorrect (reality evaluations) and as are involved in judgments of personal success or failure (achievement evaluations)" [p. 293].

In contrast to these rationalist approaches, which lead to the assumption of universal principles of moral judgment, a growing body of recent research is apparently motivated by a descriptive approach referring to moral intuition (Haidt, 2001; Saltzstein & Kasachkoff, 2004), to heuristics, ideological narratives (Haidt, Graham & Joseph, 2009), the blinding function of morality (Haidt, 2013), and the affective influence of sacred values (Graham & Haidt, 2012).

[1] http://plato.stanford.edu/entries/morality-definition/#toc

Rather than implying universals, this intuitionist and affective approach implies variation between cultures, groups and ideologies, pluralism in moral values (Berlin, 1998), and distinct relations between moral rules and (left vs. right) political orientations (Haidt &T Graham, 2009; Van Leeuwen & Park, 2009). Moral judgment is strongly qualified by cultural values, group conventions, religions, or political and epistemic authorities. As Haidt and Graham (2009) put it, the domain of morality cannot be restricted to harm and fairness, its classical issues. It must be broadened to include community (ingroup), authority, and sacredness (Durkheim, 1925/1973) as distinct sources of bias and irrationality. Haidt (2013) raises this new research trend to the level of a prescriptive advice: "I offer three principles that I think should characterize moral psychology in the twenty-first century: (1) Intuitions come first, strategic reasoning second, (2) There's more to morality than harm and fairness and (3) Morality binds and blinds" [p. 281].

Ethical Norms and Moral Attribution Among Scientists

The present article is neither concerned with moral dilemmas (Nichols & Mallon, 2006), the currently most prominent experimental paradigm, nor does it directly speak to harm, justice, ingroup, authority, and sacredness, the key concepts in of intuitionist approach. It is rather concerned with ethical and moral evaluation of scientific practices – an issue that has recently become the focus of a widely attended, self-critical and sometimes self-depreciating debate. Although this topic may appear a bit self-centered and detached from more common areas of morality research (e.g., existential moral dilemmas), it has considerable consequences for the public image of science, for funding schemes, and for the future development of scientific methods and practices. Above all, this topic affords an reality test of whether the two prominent approaches can account for an up-to-date moral debate.

Compliance as an end in and of itself. It will soon be apparent that this current discourse on ethical violations, immoral transgressions and questionable practices in science neither meets

the criteria of mature stages of moral judgment according to Piaget, Kohlberg or Kelley, of reciprocity, and mastery, nor can it be understood within the intuitionist framework. Rather, this debate appears to be mainly driven by a well-known heuristic that has been almost forgotten in the pertinent research, namely, the impact of conformity and compliance with existing rules conceived as a normative end in and of itself, independent of its rationality and utility and often detached from real authority, solidarity, or sacred value. Regardless of the lessons provided in Hannah Arendt's (1963) *Eichmann in Jerusalem: a report on the banality of evil* and in Asch's (1956) or in Milgram's (1963) memorable demonstrations of conformity, the mere compliance with given rules, instructions or conventions continues to dominate the discourse on good practices in science. Pertinent evidence for this claim will be provided in the next section.

It may indeed be no exaggeration to say that compliance has hardly ever enjoyed the same popularity as in these days, given virtually unlimited access to personal data and technologies for the monitoring and storage of everyday compliance. Almost every large or middle-size company has its own compliance department to monitor and control ethical behavior, well below the threshold of legal transgression. Conformity with short-sighted norms is given more attention than compatibility with higher-order legal principles and human rights. Good or bad intentions are hardly considered when sanctioning non-compliant people (like Edward Snowden). What is politically correct and conformist not only determines what is feasible in politics, but even in science (using citation indices as a conformist diagnostic criterion) and in interpersonal behavior. Even the social status of young children depends on their compliance with brands names, popular idols, jargon, and behavioral routines. Asch (1956) already found that experimental participants who provide non-conformist responses become disliked and devalued. And, modern experiments on group decision making continue to demonstrate that dissenters who dare to utter unshared information are disliked (cf. Wittenbaum, Hubbell & Zuckerman, 1999).

Moral reasoning among scientists. Note that uncritical compliance with haphazard standards and instructions, even in the absence of strong authorities or forces of ingroup solidarity, can be neither rational nor can it be due to sacred values and evolutionary or cultural foundations or morality. An interesting and thought-provoking question is, therefore, why rational scientists – especially those who did take their lessons from Asch, Arendt, and Milgram – should exhibit mere compliance rather than living up to a mature and rational level of moral judgment. Are the ethical norms and moral values that characterize the behavior of scientists not obliged to broadminded social ideals and interests, without regard for the presence of interest holders, and without regard to the individual's own interest (cf. Kelley, 1971, p. 294)? Should the moral evaluations and attributions of scientists be determined by the same shallow rules of compliance that have been shown to characterize social behavior at modest levels of reflection?

In the remainder of this article, I first provide a sketch of the recent debate on violations of ethical and moral norms and rules of good scientific practices. I will then argue that this debate is neither characterized by Kelley's (1971) notion that judgments of reality (correct vs. incorrect) and achievement (success vs. failure) underlie moral evaluation, nor does it reflect the affective impact of actual harm, injustice, ingroup interests, real authority, and violations of any sacred norms. It rather seems to reflect the Zeitgeist of a compliance-type of "good-boy" morality (Kohlberg, 1963) that does not appear to be sensitive to higher levels of attribution (see Table 1) or even intuitive judgments of honesty, fairness, and hypocrisy.

Major Topics of the Current Debate on Ethical and Moral Conduct in Science

What major topics or types of morally questionable behaviors that have been the focus of so many recent articles?

Plain fraud. First of all, there have been a few cases of actual fraud, in which behavioral scientists finally admitted to have committed intentional and systematic data fabrication. These

are severe transgressions that not only give unfair advantages to fraudsters, who surreptitiously profit from major publications in high-reputation journals, but also serve to the undermine the public image and the trust in scientific work. Everybody will agree that the scientific community has a vital interest in diagnosing, understanding, and sanctioning data fabrication.

Plagiarism and violations of authorship rights. The same holds for obvious cases of plagiarism, which have become a popular issue in Germany, due to the revelation that prominent politicians' dissertation had been stolen. However, despite the negative mass-media influence of these prominent affairs on the public image of academic institutions, plagiarism does not appear to play much of a role in current psychology. On the contrary, the scientific community appears to be quite insensitive to plagiarism and disinterested in protecting authorship as a valuable good. We never really care about whether every individual in a long list of authors really made a substantial contribution, or whether the senior author in the last position contributed more than equipment and resources. We not even care much about a useful criterion for authorship. Also, the recent claim for unlimited access to data and research tools reflects wide compliance with the transparency claim, but little sensitivity for protecting authorship in the creative science process, obviously because there is no majority or proponent of authorship or intellectual origin.

Harmful consequences of applied research. There is also a conspicuous lack of interest in ethical problems associated with applied psychology, comparable to the ethical debates about genetically manipulated crops, therapeutic use of stem-cells, or animals killed in biology and life science. Although applied psychology is much more likely to cause manifest harm, costs, and personal injustice than fundamental research (due to inappropriate diagnosis, treatment, or faulty methodology), there is nothing comparable in applied psychology, outside the lab, to the ongoing ethics debate revolving around practices within the lab. Lawyers, journalists, and peer researchers do not care much about payment for ineffective or inappropriate psychotherapy, the failure of

older therapists to participate in further vocational training, malpractice and methodological mistakes of expert witnesses leading to wrong legal decisions, discrimination and bias in personnel selection, vested interested and subtle forms of corruption, unwarranted use of questionable survey data, or irresponsible publications of scientifically dubious findings or unwarranted interpretations – an issue to be taken up soon.

Attribution of fraud. If my perception is correct, severe forms of fraud, plagiarism, or corruption do not appear to play strong roles in psychological research, and the ethical debate largely excludes those applied areas, where they are most likely because much is at stake. Still, even when fraud exists in psychological research, one might ask whether an attributional analysis using Kelley's (1971) criteria would justify an association of fraud with psychology. Using Heider's (1958) levels of attribution scale – borrowed from Shaw and Sulzer (1964) and summarized in Table 1 – it would be interesting to see whether the attribution of fraud in science exceeds even the most primitive level of global associations (of fraud with psychology).

On one hand, granting a rate of non-zero deception in all domains of life, a logical truism is that deliberate sampling of negative cases will always discover a few existence proofs of fraud. If so, the very existence of a highly selective non-zero sample does not tell us anything about the prevalence of fraud. On the other hand, only a small minority of documented fraud cases came from within psychology; the prevalence of fraud is much higher in life sciences (Stroebe, Postmes & Spears, 2012). An even more telling comparison could be based on a systematic assessment of fraud in all areas of cultural life, such as politics, banking, business, journalism, legal affairs, sports, or close relationships. From such a broader perspective, science may well appear to represent an idyllic place of relatively high rates of mutual trust and honesty.

Whether this is true or not, any reasonable moral judgment ought to try going beyond the most primitive level of global association, also known as the fundamental attribution error

(Tetlock, 1985), and instead try to take external circumstances and constraints into account. Crucial to attaining a more mature level of attribution is the mastery of the tradeoff between intentionality and effect strength (Piaget, 1932). While young children believe that a large (expensive) window broken by a clumsy football kick is more severe than a small window broken by an intentional kick, the moral judgments of older children give more weight to intention than to the effect. More generally, an advanced stage of morality is evident in shifting weight from the (visible) effect to the (latent) cause.

Questionable research practices. The failure to reach such an advanced level of moral reasoning is apparent in the malicious discussion of so-called questionable research practices that, unlike severe deception, appear to be quite common in behavioral science. Thus, the greatest part of the discourse revolves around such behaviors as: not reporting all obtained findings in a published paper, selectively reporting studies that worked, or claiming to have predicted unexpected results. In a frequently cited paper by, John, Loewenstein and Prelec (2012) came to conclude, based on a survey among active scientists, that such "questionable practices may constitute the prevailing research norm" [p. 524] and that their prevalence "raises questions about the credibility of research findings and threatens research integrity" [p. 531]. These conclusions are so far-reaching and threatening for the scientific community that a more informed assessment of the underlying evidence is in place.

First of all, it is unfortunate and unjustified that the discussion of these questionable research practices is often immediately linked to the recent cases of deliberate fraud, which are not comparable in terms of severity and underlying motives. There is no reason to assume that questionable practices of the aforementioned kind cause fraud. If somebody is hard-boiled enough to engage in plain cheating, he or she does not have to exhibit these subtle biases in self-presentation and selective reporting. Placing questionable practices and fraud in close context

only serves to further enhance the global association of psychology with a diffuse meaning of immoral behavior.

More importantly, a closer analysis of the questionable research practices reveals that most of them may be attributed to external circumstances and attributes of the scientific system rather than internal dispositions of the individual researcher. Selectively reporting studies that have worked may simply reflect the truism that editors and reviewers do not allow for the publication of results that did not work. The very fact that authors do not continue to try it again and again simply means that they are anticipating the publication bias that exists as a system constraint. Interpreting these cases as "researcher practices" may thus be a misnomer, an ordinary example of the fundamental attribution error, that is, to mistake external conditions for personal dispositions. Similarly, not reporting all the results that were obtained in a study and in the subsequent data analysis (often including simulations and sub-analyses motivated by diverse ideas emergent in a dialectic process) may be neither unusual nor motivated by bad intentions. People who endorse not reporting all they have done may be just honest, sincere, and maybe even a bit proud of how carefully and richly they analyze their studies.

In a slightly different vein, raising the impression that all unexpected results were predictable from the beginning may reflect a normal (unintended) hindsight illusion (Fischhoff, 1975). In the light of the evidence obtained in a study, theoretically minded researchers may spontaneously engage in reconstructive attempts to give theoretical meaning to those results, and this may indeed create the hindsight illusion that "I knew it all along". Or, hindsight interpretations may simply reflect the impact of reviewers or editors who force authors to articulate an account of all reported findings. In any case, the behavior may not originate in the researcher's intention to raise too positive an impression of his or her work. The actual motive may indeed be prosocial (helping readers to understand the research) or compliant (adhering to an

implicitly learned writing norm) or self-deceptive (not remembering the original prediction). Note also that the epistemological status of a hypothesis, whether and when it was adopted by the researcher, is unlikely to have any substantial effect on the data analysis (which is typically determined by the design), on the assessment of internal and external validity, on the chances of the study to be included in reviews and meta-analyses, or on any other theoretical and practical consequence of the study. If so, "admitting" that one has not fully anticipated a hypothesis is rather just a humble act of compliance, like an arbitrary item in a lie detection test, or a lying subscale of a personality inventory.

Minor deceptions. To be sure, the John et al. (2012) survey also included some other, less equivocal practices that come closer to dishonest behaviors with unwanted consequences for the scientific process, such as rounding-up *p*-values or claiming that results are unaffected by demographic variables when this was actually not tested. Let us assume that such behaviors actually represent minor cases of dishonesty or "white lies", apparently aimed at getting a paper past editors or reviewers who may not publish a study because of an unexpected gender effect or because a significance test resulted in $p = .052$ rather than $p = .049$. Even when one (like the present author) does not want to excuse plain lying even on minor details, a fair assessment of morality cannot fully exclude the role of external causes. Is such hypocritical, petty-minded behavior not to a notable degree reflective of features in the scientific system that evoke hypocrisy (in adaptive players) and therefore call for external attributions?

Compliance as a Consequence of Single-Sided Morality

The current discussion of questionable research practices is not only superficial and unsophisticated from an attributional point of view; it is also narrow-minded in terms of the behaviors under focus. The greatest part of the published debate is concerned with practices that undermine the assumptions of statistical hypothesis testing, increasing the danger of α-errors in

particular (Simmons, Nelson, & Simonsohn, 2011). Stopping data collection after obtaining the desired results or excluding outliers after looking at the impact of doing so are unwanted behaviors because they violate the assumption of random sampling and stochastic independence in hypothesis testing. As a consequence, the nominal α in significance testing is no longer the actual error probability that an effect observed in a sample will be obtained under the null hypothesis. Not complying with these rules enforced by inference statistics is considered a sin that undermines valid and replicable science. All other facets of methodology related to theorizing, logic of science, terminological precision, research design, and modelling are largely ignored and virtually never considered morally relevant.

Compliance with the demon of statistical-significance testing. Let us illustrate this hypocrisy, consider the recent research on the enhanced memory for faces associated with the cheater detection motive, a topic of great interest in evolutionary psychology. For instance, in a study by Buchner, Bell, Mehl, and Musch (2009), participants were presented with a series of faces along with text passages that either referred to cheater detection ("Cheater K. S. is a secondhand-car dealer. Regularly, he sells restored crash cars as supposedly accident-free and conceals serious defects from the customers") or not ("Cooperator N. G. is a mechanic. He is always eager to provide spare parts as cheap as possible for his clients and to fulfill his jobs efficiently"). Significance tests clearly reveal that in a subsequent recognition test, memory for faces associated with cheating is clearly superior to memory for faces associated with cooperation or neutral themes. This well-replicated finding is consistent with the notion that cheater-detection and social exchange motives are evolutionarily significant and cognitive abilities as old as the time of hunters and gatherers (Cosmides & Tooby, 2004).

Could the late exclusion of a few outliers (who have apparently not understood the instructions) or a questionable stopping rule (ceasing to sample participants when the available

evidence already demonstrates the expected results) greatly reduce the scientific value of this kind of research? An informed answer clearly depends on an analysis of the alternative condition, namely, the same research in the absence of those statistical malpractices. Can strict adherence to the rules of significance testing be presupposed to result in unbiased and valid inferences about the cheater-detection hypothesis?

Some closer inspection reveals that under most study conditions the strong assumptions of significance testing are unlikely to be met anyway. In a recognition test involving dozens of faces, it is hardly justified to assume that memory measures for different faces are stochastically independent, that discriminability and response bias are stable over the entire test time, that all assumptions about scaling and measurement resolution are met, or that the sample of participants can be considered truly random. So, it would only be honest to admit that the nominal α is virtually never the true α and, even more directly, a significance test is never more than a crude heuristic to judge the viability of a hypothesis, not a sound basis for inferring the validity of the hypothesis. Again, honesty and morality requires rational reasoning: A significant finding only means that $p(\text{obtained effect} \mid H_0)$ is very low ($< \alpha$) and that there is no rationale for making inferences about $p(H_0 \mid \text{obtained effect})$ or $p(H_1 \mid \text{obtained effect})$. Let me quickly add that this logical restriction is also not overcome when significance testing relies on Bayesian (rather than Fisherian) tools. The strong dependence of Bayesian statistics on prior odds, $p(H_1)/p(H_0)$, highlights the fact that the posterior odds (i.e., that H_1 rather than H_0 is true) depends crucially on other factors than the statistical properties of the obtained effect in a sample.

Provided some more modest consensus can be reached about the omnipotence of statistics, the question is why the morality debates centers on prescriptions derived from statistics. The most reasonable answer that comes to mind – namely that precision is a major asset in science – is unwarranted when no precise inference about $p(H_1 \mid \text{obtained effect})$ is possible

anyway. On the contrary, strict adherence to statistical rules of the game might foster an illusion of confidence in the validity of findings that ought to be regarded as rather local in value. Alternatively, one might contend that discipline and commitment to rules of the game are fundamental assets, symbolic symptoms of trustworthiness or of a general disposition to be honest and committed to the scientific community. In my opinion, though, a more convincing answer would have to point out that blind compliance with unreflected rules is at work. Any other account could hardly explain why so many other sources of invalidity, bad research design, measurement and sampling error, and sloppy interpretation are not also treated as dangerous for scientific precision and integrity and as morally relevant.

Conspicuous insensitivity to non-statistical dangers. A more open-minded valuation of dangers that might undermine the trust in and reliance on science has to realize, first of all, that the randomized sampling of participants represents but one of a variety of sampling filters (Fiedler, 2011). Inferences about the validity of the cheater-detection hypothesis not only depend on the size and the allegedly pure randomness of the sample of participants. They also depend on the sampling of stimuli (faces used), treatments (text passages referring to cheating or other behaviors), levels manipulated on independent variables (in a typical fixed-effects design), indices used to measure the dependent variable, different wordings of instructions, task settings (including time delays; attentional constraints; list length; etc.), or psychologically relevant boundary conditions of recognition performance (e.g., mood, regulatory focus etc.). The principle of representative sampling (Brunswik, 1955; Dhami, Hertwig & Hoffrage, 2004) calls for study designs that treat all these aspects as random factors, rather than fixed-effect factors restricted to two or very few arbitrarily selected study conditions (cf. Wells & Windschitl, 1999). However, while perfect compliance with the norm of random sampling of participants is considered crucial for good practices in science, the failure to render a study representative all these other respects is

not part of the arbitrary behavioral code. Not checking on the degree to which face memory is peculiar to the specific text passages presented with the faces, to the strength of the cheating appeals, to demand effects inherent in specific instructions, or to boundary conditions like incidental versus intentional memory settings would not be considered questionable a practice.

A provocative paper by Vul, Karris, Winkielman, and Pashler (2009) that focused on inflated correlations ("voodoo correlations") in neuroscience is telling about the serious validity problem that results from selective sampling of measurement points ("voxels"). The authors had pointed out that correlations as high as $r = .70$ or more between brain measures and behavioral measures of traits or performance might be due to the fact that such correlations are often based on a highly selective subset of those voxels (out of 150000 or so) that bear the strongest relation to the behavioral criterion. If so, this might not only undermine the validity of neuro-science but actually border on invalidity of a morally significant kind. However, neuroscientists were quick to deny that 150000 degrees of freedom are reduced like that. They rather pointed out that the selection of voxels in state-of-the-art neuro-research is based on regions of interest (ROIs), or brain areas that have been shown previously, or in pilot studies, to be relevant to the explanation of behavioral correlates. In other words, to exculpate a researcher from any accusation, the reframing of a voxel selection stage as a pilot study is sufficient. If researchers use early participants (framed as a pilot study) to select voxels defining a ROI empirically, this is deemed to be logically (and morally) different from a selection framed as initial stage of a single main study. The example highlights the weakness of the underlying moral code that is used to classify behaviors as questionable or not, and it highlights again the need to anchor moral valuation in a refined attributional analysis, beyond global associations and moral intuition.

Failure to engage in deep and responsible theorizing. The need to engage in moral reasoning proper is most evident when it comes to careless (and often self-serving or self-

deceptive) theorizing. Consider again the cheater-detection example. If the failure to stop sampling according to a pre-determined rule and the post-hoc exclusion of invalid participants constitute morally questionable practices, then how serious is a premature theoretical explanation of a finding for which there is no logical foundation? Indeed, Buchner and Bell (2012) first succeeded in gathering strong evidence for the hypothesis *If cheater then enhanced memory* across several replication experiments, using large samples and apparently clean sampling methods. However, reminiscent of Peter Wason's (1960) lesson on conditional reasoning – that *if p, then q* does not imply that only *p* affects *q* – they then ran another experiment to see whether cheater detection is really the crucial causal factor. Granting that the text passages used to induce cheater detection could also serve to induce cheater-independent negative meaning, they also associated faces with other negative meanings and, not too surprisingly, they obtained a similar increase in memory performance.

I anticipate that hardly anybody will blame researchers who fail to take logical rules into account – even though Peter Wason's lesson is as popular as the rules of significance testing. However, I ask myself why careless logical reasoning is less relevant for good conduct than complying with statistical norms. The only plausible answer that comes to my mind, related to distinction of compliance and conversion (Moscovici & Personnaz, 1991), is that statistical norms are a matter of compliance whereas logical reasoning calls for critical, emancipated reasoning – lying outside the realm of common interpretations of good scientific practices.

Deeper reflection and analysis shows that careless theoretical reasoning – which is of course not hap-hazard but typically favoring researchers' beloved hypotheses – is a wide-spread phenomenon (Fiedler, Kutzner & Krueger, 2012). Wason (1960) had shown that when trying to identify the rule underlying a sequence like 2, 4, 8 people quickly come up with overly specified rules (such as 2^N) and restrict their hypothesis test to checking the predictions of the selected rule,

finding out that, say, 16, 32 and 64 also provide positive examples of the rule. Using such non-Popperian strategies, they never find out that the correct rule might not be 2^N but some less specific rule that includes 2^N as a special case. For instance, they fail to find out that the actual rule might be super-linearly increasing integers, or increasing integers, or any series of integers, or any real numbers, any numbers, or any set of symbols (whether number or not).

By analogy, researchers – even when their work is published in the best journals – interpret that impact of exposure to a funeral or a mortality-related essay topic on conservative behavior as an impact of mortality salience (Greenberg, Solomon & Pyszczynski, 1997). They hardly test whether the manipulation affects some super-ordinate construct that includes mortality as a special case. Rather than mortality, it might reflect the priming of existential values (implying similar effects for birth as for mortality), or self-referent affect, or simply priming of incompleteness (implying similar effects for mortality-unrelated incompleteness; Wicklund & Braun, 1987). In a similar vein, the manipulation of exposure frequency is presupposed to causally induce fluency, rather than other, more general aspects of density in memory. Or, returning to Buchner and Bell (2012), one need not take it for granted that vignettes manipulate cheater detection, rather than other negative meaning or other non-negative affective meaning.

The naïve believe in the validity of one preferred theoretical account, and the failure to check for a whole variety of alternative causal models of the same findings, is particularly evident in the epidemic use of mediation analysis (Fiedler, Schott & Meiser, 2011), which is general considered a gold standard of good practice and strong science worthy of imitation. Closer inspection shows that the vast majority of all published research that uses mediation analysis only ran a statistical test that focused on one favorite mediator. It rarely happens that researchers engage in comparative tests of several mediator candidates, and researchers virtually never test for alternative causal models but mediation (cf. Danner, Hagemann, & Fiedler, 2015).

Thus, rather than mediating the influence of X on Y according to a causal mediation model X → Z → Y, the third variable Z might be a covariate of Y in a common-cause model X → X,Z, or all three variables might just be drawn at random from a set of homogeneously correlated measures of the same syndrome. Simulation results make it crystal-clear that statistical test of Z as a mediator cannot discriminate between these different causal models (Fiedler et al., 2011). Statistical tests will also be often significant when the actual causal model is not different from mediation. However, whereas careless mistakes in conducting and reporting a statistical mediation test would be a candidate for questionable practice, leading to the downgrading or even disqualification of scientists who commit the mistake, the widely shared practice of drawing premature and often wrong inferences from highly selective, single-eyed mediation tests is hardly recognized as problematic. Eventually, such misleading, hypothesis-confirming, and self-serving mediation analysis would be considered better than not running a mediation analysis. After all, the latter involves compliance with a majority habit, and less emancipation is required to diagnose a formal statistical mistake than to reason critically about alternative causal models.

More generally, the new interest in quantitative model fitting is quite in line with a compliance-oriented valuation system. By conducting mathematical model tests, researchers subscribe symbolically to precision and strictness as laudable norms of scientific research. In contrast, critical questions about whether metric data qualities assumed in quantitative models are met, whether model fit entails capitalizing on chance, and the insight that a fitting model need not underlie the data to be explained (Roberts & Pashler, 2000) would run against the easily executed compliance rule. Such critical counter-arguments are therefore not appreciated as symptom of good scientific practices and of the researcher's honesty and responsibility.

Reciprocity, Equality, and Generally Binding Moral Rules

So far, I have argued that the behavioral scientists' ethical norms and behavioral codes do not live up to higher levels of attribution that would allow them to go beyond the fundamental attribution bias (i.e., blaming the researcher rather than the system) and beyond the compliance with superficial rules. The aim of the present section is to demonstrate that even much more primitive moral rules of fairness, such as the age-old reciprocity principle or the equal-rights-and-equal-duty rule, are not visible in the current debate. Rather, there appears to be an asymmetric allocation of roles in the science game, with some people being "prosecutors" or monitor of ethical standard and other taking the role of "defendants" or targets of evaluation. Good practices are expected of and controlled in authors who want their manuscripts to be published in journals, but good practices are hardly ever assessed in reviewers and editors, the major determinants in the publication system. Similarly, good practices and minimal standards are applied to the grant-proposals writers, for PhD students, and for original researchers but hardly for reviewers of grant proposals, doctoral advisors, or scholars who blame others in published articles of questionable practices. It appears that those agents who take the offensive role of referees, evaluators, and censors need not be afraid that they will themselves be evaluated according to the same rules of good conduct that they evaluate in the targets or patients of the unidirectional compliance game.

Thus, John et al.'s (2012) widely cited survey study entailed a grave and generalized accusation that questionable research practices of the type discussed above have become a prevailing research norm so that the credibility of research findings and integrity are questioned. However, granting that John et al. are driven by a moral (compliance) norm, there is little interest in the question of whether their survey itself lives up to standards of good science. We already discussed that no deliberate attempt was made to avoid, or to diagnose, misunderstandings of questionable practices? Moreover, other rules of good survey research related to the logic of

conversation were not attended to. For instance, the only way for a respondent to appear innocent would have been to respond "No" to all items, which cannot be expected in a survey.

However, the most serious shortcoming is that John et al. assessed the proportion of people who ever committed a behavior (e.g., ever stopped sampling once the desired results were obtained) at least once in their life, and the resulting proportions were then treated like evidence on the prevalence of these behaviors. This is of course a category mistake because researchers conduct many hypothesis tests in their lives, and the prevalence of the behavior is some multiplicative function of the proportion of scientists with a non-zero rate times the average repetition rate across all studies (cf. Fiedler & Schwarz, 2015). Nobody would come to doubt that the prevalence of lying is by magnitudes lower than the proportion of people who ever told a lie. So why should we believe that the proportion of scientists who ever committed a behavior equals the behavior's prevalence? What gives us the right to believe that any sound inference from the survey to the behavioral prevalence is possible at all?

Indeed, our own experience with a critical test of the John et al. (2012) data (leading to strongly divergent results) is that editors and reviewers do not appreciate any attempt to blame prosecutors. Without strong arguments against the validity of our critique (Fiedler & Schwarz, 2015), they continue to believe that John et al. were driven by a laudable motive and reviewers feel intuitively that their pessimistic prevalence estimates are probably closer to the truth than the more optimistic estimates obtained in a survey with different measures for prevalence and non-zero proportions. They also don't mind that their own behavior prevents a potentially very informative paper from seeing the light of publication, although they seem to agree that not reporting unwanted findings is a bad practice. Crucial to understanding the contradiction is of course the fact that the John et al. article fits neatly into what is considered politically correct and therefore likely to become a compliance norm.

In a similar vein, Simonsohn, Nelson, and Simmons (2014) refer to "*p*-hacking" as an explanation of the conspicuously high rate of α values slightly lower than .05 in significance tests of published articles. The action verb "*p*-hacking" entails an internal attribution to researchers' intentional, deliberate actions, although an obvious external attribution would be to understand the peak at slightly below $\alpha = .05$ as a reflection of a filter in the publication system. Even though I sensitized them to the surplus meaning of the verb and its implications, they decided to continue using the term *p*-hacking. Obviously, if one is on the appropriate side of a compliance game, one need not refrain from unwarranted depreciation of peer researchers.

A slightly different but related example concerns the unfortunate publication of Bem's (2011) parapsychological paper on precognition – which also contributed a lot to the current Zeitgeist of questioning scientific standards. Obviously, the decision to publish this paper was driven by compliance with the norm to treat all submissions according to the same (mainly statistical) rules and not to be prejudiced against research in parapsychology (Judd & Gawronski, 2011). Afterwards, the published critique of Bem's work was largely restricted to issues of appropriate significance testing – another compliance domain. Other serious problems with the logic of science and the validity of Bem's research, which I pointed out repeatedly as a reviewer from the beginning, were hardly ever noted and only published in a low-publicity journal (cf. Fiedler & Krueger, 2013). In leading empirical journals there is no room and little interest in enlightening debates if they do not refer to statistical significance testing.

From all perspectives, as editor reviewer, and as an author who submits his own papers to leading journals, I have witnessed reviewers deliberately arguing against the publication of research that works against their own interests or previous results. If this occurs – and it does occur regularly when reviewers are really expert in a field – my impression is not that reviewers have much to lose in ethical reputation if they engage in clearly one-sided critique and obvious

attempts to turn a paper down. They possess an evaluator role. Similarly, it is relatively easy for reviewers to allude to the alleged fact that some evidence is not new, without indicating a reference, or simply to express that they do not like what researchers have done. My impression is that editors who act as arbiter rather than simply counting and following reviewer votes are the exception rather than the role. Although the impact of reviewers and especially of editors on publication decisions is much stronger than the impact of the authors' practrices, the former are unlikely to become the target of moral valuation. The implicit norm is apparently that editors publish an article if they want to publish it and that reviewers are doing valuable honorable work that does not deserve to be judged morally. The only role that is weak enough to be assigned a patient part in the compliance game is the author, who has to provide signed declarations of good conduct and to worry about being responsible for published data and interpretations. There is little need to for the more powerful roles in the game to develop guilty feelings, although their impact on the publication output is strongest and most direct.

Conclusions: Can Social Psychology Account for Morality in Science

Thus, it seems obvious that the reality of a recent moral debate among in and around the scientific community can be neither explained in terms of the old rationalist approaches to morality nor in terms of modern intuitionist approaches. On one hand, the manner in which questionable practices are defined and in which scientists are held responsible for any effects of their actions, regardless of intention and foreseeability and regardless of how many people engage in similar behavior, would be at best classified as Level II in Heider's scale (cf. Table 1). The underlying moral principles are far away from Levels V or IV, which ought to be reached according to Piaget (1932) and Heider (1958) by moral maturation alone. In fact, I do believe that many scientists would defend the rigid Level-II rules as an asset, a precondition for objectivity in good science. But note that Piaget's (1932) use of the term "objective responsibility" for Level II

was meant to denote unsophisticated moral judgments that are insensitive to motives and intentions, rather than objectivity in checking violations of compliance rules. Apparently, the current Zeitgeist is more interested in establishing the latter meaning of objectivity than in overcoming the former meaning.

On the other hand, scientists' moral judgments are not only insensitive to higher levels of attribution, lying outside the domain of rationalist theories in Piaget's and Kant's tradition. They also lie outside the domain of the modern intuitionist approach, with focuses on affective heuristics and sacred values, which do not appear to motivate the current debate. With respect to the key concepts harm, justice, ingroup, authority, and sacredness (Haidt, 2013), it has to be noted that in this debate (a) no cost-benefit analysis is conducted to assess harm; (b) no fairness rules of justice are applied to everybody; (c) no ingroup-serving bias prevents whistleblowers from blaming their own ingroup members; (d) no natural authority is apparent behind the debate; and (e) there is nothing sacred in the most widely respected statistical norms.

My motivation here is not to criticize my peers or my scientific community for a moral attribution style that others have considered immature. My motive rather is to point out that the most prominent psychological approaches to morality may not be applicable to real-world manifestations of moral evaluation. I hasten to add that focusing on morality in other parts of real life – such as the conduct of politicians or journalists, democratic rights and duties, commercial business rule, or faithfulness in close relationships – would probably lead to the same conclusion: Neither rational rules (cost-benefit analysis, reciprocity, consideration of norm distributions) nor phylogenetically inherited moral heuristics capture the essence of moral judgment under realistic conditions (see also the critique by Saltzstein & Kasachkoff, 2004).

The one approach that in my view comes closest to understanding the motives, the monitoring mechanisms and the moral judgment rules in contemporary science can be found in

the seminal writings of Hannah Ahrendt (1963), Solomon Asch (1956), Stanley Milgram (1963), and Serge Moscovici on uncritical and strict reliance on compliance rules, contrasted against rational and intellectually advanced rules of compliance, argumentation, and mature attribution.

I believe that content validity for broadly assessed morality issues in real life should not be neglected in social psychological research. I am not taking for granted that the analysis I have outlined in the present chapter is the only viable perspective. Proponents of intuitionist research may have other reality domains in mind, which I have overlooked, and maybe some scientists involved in the pursuit of good scientific practices can convince me that the underlying moral rules are more sophisticated than I could see. For such counter-arguments to be really convincing, though, they would have to come up with real evidence that goes beyond experiments in which moral dilemmas are described in vignettes that enforce sacred values and emotional instincts that are rarely invoked in everyday reality.

To the extent, however, that moral judgment and action in real life is indeed driven by compliance with (often arbitrary but objectively applicable) norms of conduct, this has not only implications for a refined theory of moral judgment. It also has obvious practical implications. Scientists and practitioners – in politics, law, economy, education, and therapy – have to reflect on whether morality is of practical value and, if so, to analyze the relation between moral means and moral ends. Are there good reasons to assume that objectively applicable compliance means foster the attainment of such moral ends as validity in scientific, reducing inequality in the global world, affirming human rights and personal dignity, and fairness in sports and courtrooms? Or might a comprehensive analysis – which has to be both moral and scientific – show that it is functional and worthwhile in the long run to strive for higher levels of moral attribution, beyond mere compliance and sacred values?

References

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

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*(9), 1-70. doi:10.1037/h0093718

Bem, D. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*, 407-425. Doi: 10.1037/a0021524

Berlin, I. (1998). My intellectual path. *New York Review of Books,* 14 May.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*(3), 193-217. doi:10.1037/h0047470

Chabris, C. F., Hebert, B. M., Benjamin, D. J., Beauchamp, J., Cesarini, D., van der Loos, M., & ... Laibson, D. (2012). Most reported genetic associations with general intelligence are probably false positives. *Psychological Science*, *23*(11), 1314-1323. doi:10.1177/0956797611435528

Cialdini, R. B. (2007). Descriptive social norms as underappreciated sources of social control. *Psychometrika*, *72*(2), 263-268. doi:10.1007/s11336-006-1560-6

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. doi:10.1037/0033-2909.130.6.959

Durkheim, E. (1925, 1973). *Moral education* (E.Wilson & H.Schnurer,. Trans.) New York: The Free Press. (Original work published in 1925).

Fiedler, K. (2011). Voodoo correlations are everywhere—Not only in neuroscience. *Perspectives on Psychological Science*, *6*(2), 163-171. doi:10.1177/1745691611400237

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from α-error control to validity proper: Problems with a short-sighted false-positive debate. Perspectives on Psychological Science, 7(6), 661-669.

Greenberg, J., Solomon, S., & Pyszczynski, T. (1997). Terror management theory of self-esteem and cultural worldviews: Empirical assessments and conceptual refinements. In M. P.Zanna (Ed.) , *Advances in experimental social psychology* (*Vol. 29*, pp. 61– 139). San Diego, CA: Academic Press.

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, *1*(3), 288-299. doi:10.1037/0096-1523.1.3.288

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*, 814–834.

Haidt, J. (2013). Moral psychology for the twenty-first century. *Journal of Moral Education*, *42*(3), 281-297. doi:10.1080/03057240.2013.817327

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, *20*(1), 19. doi:10.1007/s11211-007-0034-z

Haidt, J., & Graham, J. (2009). Planet of the Durkheimians, where community, authority, and sacredness are foundations of morality. In J. T. Jost, A. C. Kay, H. Thorisdottir (Eds.) , *Social and psychological bases of ideology and system justification* (pp. 371-401). New York, NY, US: Oxford University Press. doi:10.1093/acprof:oso/9780195320916.003.015

Haidt, J., Graham, J., & Joseph, C. (2009). Above and below left–right: Ideological narratives and moral foundations. *Psychological Inquiry*, *20*(2-3), 110-119. doi:10.1080/10478400903028573

Heider, F. (1958). *The psychology of interpersonal relations*. Oxford: Wiley.

Ioannidis JPA (2005) Why Most Published Research Findings Are False. PLoS Med 2(8): e124. doi:10.1371/journal.pmed.0020124

John, L.K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questinable research practices with incentives for truth telling. *Psychological Science, 23*, 524-532. Doi: 10.1177/095679611430953

Judd, C. M., & Gawronski, B. (2011). Editorial comment. *Journal of Personality and Social Psychology*, *100*(3), 406. doi:10.1037/0022789

Kant, I. (1788/1999). Critique of practical reason. In M.Gregor (Ed., Trans.). *The Cambridge edition of the works of Immanuel Kant: practical philosophy* (pp. 137-276). Cambridge, UK, Cambridge University Press.

Kelley, H. H. (1971). Moral evaluation. *American Psychologist*, *26*(3), 293-300. doi:10.1037/h0031276

Kohlberg, L. (1963). The development of children's orientations toward a moral order: I. Sequence in the development of moral thought. *Vita Humana*, *6*(1-2), 11-33.

Milgram, S. (1963). Behavioral Study of obedience. *The Journal of Abnormal and Social Psychology*, *67*(4), 371-378. doi:10.1037/h0040525

Moscovici, S., & Personnaz, B. (1991). Studies in social influence: VI. Is Lenin orange or red? Imagery and social influence. *European Journal of Social Psychology, 21*(2), 101-118. doi:10.1002/ejsp.2420210202

Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition, 100*, 530–542.

Piaget, J. (1932). *The moral judgment of the child*. Oxford, England: Harcourt, Brace.

Saltzstein, H. D., & Kasachkoff, T. (2004). Haidt's Moral Intuitionist Theory: A Psychological and Philosophical Critique. *Review of General Psychology*, *8*(4), 273-282. doi:10.1037/1089-2680.8.4.273

Shaw, M. E., & Sulzer, J. L. (1964). An empirical test of Heider's levels in attribution of responsibility. The Journal of Abnormal and Social Psychology, 69(1), 39-46. doi:10.1037/h0040051

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. doi:10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534-547. doi:10.1037/a0033242

Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science, 7*, 670-688. DOI: 10.1177/1745691612460687

Tetlock, P. E. (1985). Accountability: A social check on the fundamental attribution error. *Social Psychology Quarterly*, *48*(3), 227-236. doi:10.2307/3033683

Van Leeuwen, F., & Park, J. H. (2009). Perceptions of social dangers, moral foundations, and political orientation. *Personality and Individual Differences*, *47*(3), 169-173. doi:10.1016/j.paid.2009.02.017

Wittenbaum, G. M., Hubbell, A. P., & Zuckerman, C. (1999). Mutual enhancement: Toward an understanding of the collective preference for shared information. *Journal of Personality and Social Psychology*, *77*(5), 967-978. doi:10.1037/0022-3514.77.5.967

Table 1: Heider's (1958) levels of attribution of responsibility, using formulations borrowed from Shaw & Sulzer (1964)

Level	***Attributions at this level are characterized by the follows moral rules:***	***Roughly corresponding to Piaget's (1932):***
Level I: Global association	Person is responsible for any effect associated with the person. Mere presence might be enough.	Syncretistic, pseudocausal reasoning
Level II: Extended commission	Person is responsible for any effect of his or her actions, even when effects were unintended and could not be foreseen	Objective responsibility
Level III: Careless commission	Person is responsible for any foreseeable effect of his or her actions, even when unintended	
Level IV: Purposive commission	Person is responsible for any foreseeable and intentional effect of his or her actions	Subjective responsibility
Level V: Justified commission	Even when action effects are intentional, person is only partially responsible if most other persons would have felt and acted the same	