Persistent Evidential Discordance

Samuli Reijula and Sofia Blanco Sequeiros

May 23, 2024

Forthcoming in the *British Journal for the Philosophy of Science*

Abstract

Successful replication is a hallmark of scientific truth. Discordant evidence refers to the situation where findings from different studies of the same phenomenon do not agree. Although evidential discordance can spur scientific discovery, it also gives scientists a reason to rationally disagree and thereby compromises the formation of scientific consensus. Discordance indicates that facts about the phenomenon of interest remain unsettled and that a finding may not be reliably replicable. We single out *persistent evidential discordance* as a particularly difficult problem for the epistemology of science, and distinguish between different causes of evidential discordance – non-systematic error, noise, and bias. Unlike discordance brought about by non-systematic error or noise, persistent discordance often cannot be rationally resolved by temporarily suspending judgment and collecting more data within existing lines of inquiry. We suggest that the analysis of enriched lines of evidence (Boyd 2018) provides a useful approach to diagnosing and evaluating episodes of persistent evidential discordance. Attention to the line of evidence, which extends from raw data to an evidential claim supporting or disconfirming a hypothesis, can help researchers to locate the source of discordance between inconsistent findings. We argue that reference to metadata, information about how the data were generated and processed, can be a key step in the process of resolving normative questions of correctness, i.e., whether a line of evidence provides a legitimate answer to a particular research question. We illustrate our argument with two cases: the alleged discovery of gravitational waves in the late 1960s, and the social priming controversy in experimental psychology.

***Keywords**:* evidence, discordant evidence, line of evidence, inscription, metadata, replication

ORCID: Reijula, 0000-0001-6968-5819; Blanco Sequeiros, 0000-0001-8049-5952

1. Introduction

Insanity is doing the same thing over and over and expecting different results.

– Albert Einstein

Replication of a finding is a sign – for some, the only sign – of scientific truth (Peirce 1878; Popper 2005/1959; cf. Romero 2019). When things go smoothly, the outcomes of scientific inquiries clearly stand for or against a hypothesis or theory, and we have something like Perrin's derivation of Avogadro's number, where several experiments converge to the same finding (Meehl 1990; Hudson 2014, Ch. 4). Often things do not go smoothly, and scientists end up with evidence that is not only associated with high uncertainty, but also contradicts other, similar studies (Achinstein 2001, Ch. 1; Franklin 2002; Earp and Trafimow 2015; Boyd 2018). *Discordant evidence* refers to a situation where findings from

different studies of the same phenomenon do not agree (Stegenga 2009, 2012; Hey 2015). In the clearest case, one study gives support to the truth of hypothesis *H* and another to that of not-*H*.

Cases of discordance are well documented in the history and philosophy of science. Franklin (2002) discusses discordance in physics, Bokulich (2020b) references evidential discordance in measuring geologic time, and Ohnesorge (2022) describes a century-long disagreement in measurement outcomes on the value of Earth's ellipticity. Although evidential discordance can be viewed as a resource generating novel scientific inquiry (Ohnesorge 2021, Bokulich 2020b), from another perspective it is undeniably an epistemic problem, as it blocks evidential convergence. Discordance, particularly a type of discordance we call *persistent discordance*, means that findings cannot be reliably replicated and facts about a phenomenon of interest remain unsettled. In Section 2, we provide two examples of what we mean by persistent discordance. As a first pass, we differentiate persistent discordance from transient discordance, the latter being a short-lived discrepancy that typically dissipates with ongoing research. As our case examples show, persistent discordance is resolvable, but until then, it gives rise to reasonable scientific disagreement and can thus hinder scientific progress by preventing the scientific community from reaching a consensus.

Much of the recent literature on evidence in the philosophy of science has had little to say about evidential discordance; concordance of scientific evidence has been the typical assumption in theories of evidence.[1] Bayesian approaches have focused on the formal analysis of evidential relations, quantifying the degree of support between evidence and hypothesis, and on evidence amalgamation (e.g., Claveau, 2013; Kelly 2016; Landes, Osimani and Poellinger 2018; Lin 2022). The problem of evidential discordance in science has been addressed directly by Franklin (2002), Stegenga (2009, 2012) and Hey (2015). In addition, discussions of coherence (e.g., Bovens and Hartmann 2004), and converging measurement (e.g., Tal 2019; Ohnesorge 2021, 2022; Bokulich 2020b) address issues relevant to discordant evidence, such as the processes with which scientists are able to move from disagreeing measurement outcomes to coordinated ones.

In this article, we propose a strategy for analyzing the phenomenon of persistent evidential discordance. In Section 2, we describe two case studies that illustrate what we mean by persistent discordance: the alleged discovery of gravitational waves in the late 1960s, and the social priming controversy in experimental psychology. In Section 3, as an analytic approach to persistent evidential discordance, we distinguish between several types of evidential discordance based on its cause (error, noise, and systematic bias). In Sections 4 and 5, we turn to the notions of enriched line of evidence (Boyd 2018) and inscription (Latour 1999) to suggest a strategy for diagnosing situations involving persistent discordance. This

[1] For criticisms of traditional philosophical accounts of evidence, see Achinstein 2001; Bogen & Woodward 2005.

strategy helps pin down the precise source of evidential discordance and understand why the discordance persists. We argue that persistent evidential discordance is typically caused by systematic bias.

We show how scientific inquiry proceeds through a sequence of causal transformations from material objects and their unprocessed measurement outcomes to processed data and, finally, to propositional evidence claims that stand in an inferential relation to the hypothesis studied. Here we refer to metadata, understood as information about the data itself, including the normative commitments embodied by each inferential step in the data pipeline (the sequence of steps from data to evidence). We argue that metadata can be key in resolving epistemic questions about discordance between pieces of evidence. These questions may be descriptive (how an evidential statement was generated) or normative (which way of generating evidence is fit for answering the research question at hand).

2. Discordant Evidence in Practice: Gravitational Waves and Social Priming

The theory of general relativity predicts the existence of gravitational waves emitted by accelerating massive objects such as binary stars or black holes. Gravitational waves are weakly coupled to matter, however, and difficult to detect. Einstein himself changed his mind several times about their existence (Collins 2004 p. 29).

In 1969, after several years of experiments, physicist Joseph Weber claimed to have succeeded in detecting gravitational waves. In his experiments, Weber utilized a massive bar of aluminum alloy fitted with piezo-electric microphones. The microphones were used to detect the bar's resonation, purportedly caused by gravitational radiation. The vibrations in the 1.5m-long bar were predicted to be smaller than the diameter of an atomic nucleus. Therefore, carefully isolating the bar from unwanted sources of noise was crucial: the bar was hung by thin wires in a vacuum, and the whole setup was mounted on an acoustic stack constructed of rubber and lead sheets in order to prevent seismic events from influencing readings from the piezos. Not all noise could be eliminated, even in principle: at a temperature different from absolute zero, the thermal movement of the atoms in the bar would generate some noise. To discern gravitational wave events from random variation, oscillations of two separate bars were compared, and coincident spikes in the two time series counted. As the frequency of simultaneously occurring spikes was substantially higher than the coincidence rate predicted by chance, Weber concluded that empirical evidence for gravitational waves had been found (Collins 2004, Ch. 2, Ch. 4; Franklin 1997).

Weber's findings were considered an important breakthrough, but from the beginning, they were contested by several other physicists skeptical about the sensitivity of his experimental apparatus. Between 1970 and 1972, Weber developed his experimental setup to respond to some of the criticisms. For example, he ran his experiments simultaneously at two different sites separated by a large distance; if coincident spikes could still be detected in the two bars, this would rule out several other potential

explanations for them. By 1972, results started coming in from experiments done by other laboratories. Many of them had failed to detect signs of gravitational radiation.

In his account of the controversy, Harry Collins (1985, 2004) describes the situation as an example of *experimenter's regress*. The regress arises when a scientific disagreement concerns the existence of a novel phenomenon. Building a correctly functioning experimental setup and running the experiment requires skill, but only a successful outcome is a certain indicator that the experiment works correctly. In frontier research, however, this is the core of the disagreement: those who believe in the existence of a phenomenon think that a successful outcome means detecting the phenomenon, whereas skeptics believe that the outcome is successful when the phenomenon is not detected. At the time, no one knew how large the gravity wave signal would be, and whether the Weber bar was an appropriate device for capturing such a signal (Franklin 1997). There was no consensus even on whether there were waves to detect. In the absence of an external standard, judgments about experimenter competence and reaching the correct outcome turned into circular reasoning.

As we explain in Section 3, the controversy over gravitational waves in the 1970s serves as an example of persistent discordance. The available body of scientific evidence was discordant: some pieces of evidence confirmed the hypothesis about the existence of gravitational waves, while others disconfirmed it. Furthermore, as the controversy unfolded, the evidence generated by the different lines of inquiry by Weber and his critics did not show signs of convergence, i.e., the discordance was not transient, but persistent. It appears that only the development of new measurement technologies could move the debate forward and even then, their evidential authority in resolving the controversy would demand explanation.

Now compare the gravitational wave controversy to a more recent episode in experimental psychology, also an example of persistent evidential discordance. In 1996, John Bargh and his colleagues published experimental findings on social priming in a paper that quickly became a classic. Priming, as studied by social psychologists, is the phenomenon where environmental, observational, and perceptual stimuli activate behavioral outcomes without the subject's conscious awareness (Bargh et al. 1996). Bargh and colleagues describe three experimental setups used to study priming. In one of the experiments, the test group completed a scrambled sentence test: they constructed sentences from selected sets of words, some of which were associated with old age, such as "Florida," "bingo," "ancient" and "helpless." The control group's words held no such association. After this task the researchers, unbeknownst to the participants, measured the participants' walking speeds as they exited the laboratory. The researchers found that the participants who constructed sentences with words related to old age walked more slowly than the participants in the control group, which was interpreted as a result of priming. The researchers argued that processing words associated with old age activated a stereotypical concept of old age in the

participants' minds, which unconsciously affected their behavior (Bargh et al., 1996). This "Florida effect" was considered an important finding as it provided evidence for a novel, unintuitive explanation of the causes of human behavior; priming research suggests that many of our actions are driven not by conscious reasoning but by factors of which we are not aware of (Bargh and Chartrand 1999).

The Florida effect quickly became a celebrated finding in social psychology, but the results turned out to be difficult to replicate. In 2012, Stéphane Doyen and colleagues published a failed replication of the Florida effect study, where they provide an alternative explanation of Bargh's findings. Doyen and colleagues suggested that Bargh's results were driven by inadequate blinding: priming effects had only been observed when experimenters were aware of participants' group assignments. Over time, some critics started to wonder whether many findings on priming were spurious, that is, if the studies, often conducted with small sample sizes, confused noise for signal (cf. Harris et al. 2013).

In his response to the critics, Bargh (2012a) contested the scientific quality of Doyen's publication venue (PLoS ONE), and suggested that the failure to replicate was due to Doyen's group being "incompetent or ill-informed." Aside from such provocations, Bargh's response (2012a, b) refers to the tacit knowledge needed to perform successful replications, and to existing theory as well as published reports of replications as reasons to treat priming as a real phenomenon. Bargh suggested that small but crucial disparities between the original study and the replication attempt might account for the different outcomes. Likewise, Daniel Kahneman, who had previously written on priming but later distanced himself from Bargh's findings, did acknowledge the subtlety of priming effects and the potential impact of small changes on the results (cf. Yong 2012; Cesario 2014). In the literature on replication crisis, it is generally acknowledged that when a replication fails, it is far from clear what implications should be drawn from this failure (Earp and Trafimow 2015). To date, priming and its relevance to behavioral science remain controversial topics.

Like the controversy over the discovery of gravitational waves, the disagreement over priming manifests the experimenter's regress. Researchers from both sides of the debate questioned their adversaries' findings, methods, and competence. In the priming debate, it is unlikely that the disagreement could be resolved simply by both sides of the controversy continuing data generation by using their existing experimental setups. Lacking a crucial experiment (Earp and Trafimow 2015), the repeated experiments would fail to escape the circular dynamic of the situation.

These are just two examples of evidential discordance, a seemingly common situation in research. Sometimes discordance is intentionally generated to obscure scientific consensus and to mislead the public (Oreskes and Conway 2011), but that is not always the case. A recent study by Breznau and colleagues (2022) suggests that even when given the same data, independent truth-seeking teams conducting data analysis often end up with discordant findings. In the next sections, we turn to the analysis

of different types of evidential discordance in terms of their causes, and to enriched lines of evidence, a notion which helps diagnose sources of evidential disagreement in science.

3. Different Types of Discordance

By 1975, many members of the physics community had become convinced that Weber had not detected gravitational waves (Franklin 1997). Weber's findings were never conclusively disproven, however, and out of the six experiments that did not detect gravitational waves, five were criticized not only by Weber but also by the other critics (Collins 2004, Ch. 9). After 1975, the attention of the gravitational wave community turned to new methods, cryogenic bars and later, interferometry. Weber's lab, although marginalized, kept functioning for another two decades. In 2015, the controversy over the existence of gravitational waves was finally resolved, when the LIGO collaboration announced successful detection of a gravitational wave signal from a merger of two black holes (Abbott et al. 2016). The phenomenon of social priming has met a similar fate: confidence in the theory of priming has been one of the areas in psychology most afflicted by the replication crisis. Many psychologists still believe that priming effects exist, but effect sizes appear to be smaller and more local than originally thought, and interpersonal variation is large (see Ramscar 2016; Chivers 2019; Weingarten et al. 2016).

Once a scientific controversy has been settled and a consensus formed, the outcome often begins to appear inevitable. Therefore, for our purposes, it is useful to focus on scientific debates in the period before a consensus was reached and the argumentation process suspended (cf. Shwed and Bearman 2010). The unfolding of both debates represent cases of discordant evidence as characterized by Stegenga (2009, 2012, 2018) and Hey (2015). Stegenga (2009) distinguishes between two notions of discordance: inconsistency and incongruity. *Inconsistency* refers to the situation where similar studies, say lab experiments, produce contradicting pieces of evidence. *Incongruity* means that different kinds of studies (e.g., in the lab, in the field, or done with quasi-experimental methods) produce evidence "written in different languages," and their coherence is difficult to judge.

To clearly grasp Stegenga's distinction, we should distinguish between *data* and *evidence*. By data, we mean public records produced by measurement or experiment (Woodward 2000; Leonelli 2019, Ch. 3). Only when combined with auxiliary assumptions involved in processing and interpreting them can the data be used as evidence for a factual claim; in this sense, discordance is also a matter of interpretation. Whereas data are artifacts produced and transformed in particular contexts, the relationship between an evidential claim and a hypothesis is inferential. What makes a claim evidential is that it speaks for the truth of another claim, the hypothesis (Cartwright 2013; see Chapman and Wylie 2016, pp. 34–38). As Kuorikoski and Marchionni (2023) put it, "data simply exist, whereas evidence lives in the space of reasons."

Incongruity can be portrayed as a situation where the first scientific study concludes *P*, the second *Q*, and the third *R*, and the implications of these evidential statements on a set of competing hypotheses cannot be determined without further auxiliary assumptions. A body of evidence is inconsistent when it provides support for the hypothesis *H* but also supports some other hypotheses that are mutually exclusive with it. We regard H_1 and H_2 as mutually exclusive, for example, when they are perceived as competing explanations for an observed phenomenon. The clearest case of inconsistency is a situation where evidence from one study entails *H* and evidence from another study *not-H*.[2] In light of this distinction, both the replication controversies discussed in Section 2 are examples of inconsistency, not incongruity.

The distinction between inconsistency and incongruity suggests that the notion of discordance is not conceptually reducible to underdetermination of theory by evidence. Both inconsistency and incongruity are related to underdetermination in the sense that in neither case does the available evidence unambiguously confirm only one hypothesis out of many. In the social priming example, we cannot use all the experimental findings to discriminate between the hypotheses that priming effects do and do not exist. Each hypothesis may be supported by different pieces of evidence, but on the whole, when the total body of evidence is considered, the mutually exclusive hypotheses remain underdetermined by the available evidence. Analysis in terms of underdetermination does not help us articulate a significant difference between inconsistency and incongruity. Incongruity can be seen as a case of weak evidence according to the severity criterion given by Mayo (1996): the available evidence gives us no pressing reason to eliminate all but one hypothesis. Inconsistent evidence presents a more serious problem to scientific inference: a body of evidence from which both a claim *H* and its negation can be derived is not consistent, and from contradiction, anything follows.

Distinguishing between different causes of discordance allows us to diagnose evidential problems of different severity. There are at least three kinds of causes of evidential discordance: (non-systematic) error, noise, and bias (also called systematic error).[3] In the simplest cases, a self-contradicting body of evidence is the result of a *non-systematic error*, such as a mistyped number in a spreadsheet or a non sequitur in a chain of reasoning. Another common source for discordance is *noise*: the property of the

[2] This portrayal of inconsistency is obviously a simplification, because the coherence of information sets should not always be seen as an all-or-nothing affair. For a probabilistic treatment of the topic see, e.g., Bovens & Hartmann (2004). Furthermore, the practical implications of an inconsistency can be very different in different cases. In the cases in Section 2, what is in question is the existence of the studied phenomenon. In less severe cases, the discordance could concern the quantitative value of a variable (e.g., R_0, the rate of transmission of an infectious disease in a population, the hypothesis P standing for a particular point estimate).

[3] See Smith (2014) for a similar classification of sources of observational discrepancies and discordance in the context of measurement and metrology.

phenomenon studied may itself be unstable over time, and the data collection process typically introduces random error into the measurements. Gravitational waves observed on Earth, for example, are minute traces of major cosmological events such as a merger of two black holes or inspiraling binary neutron stars. Given the uneven distribution of such events in time, temporal variance is expected to occur in the measurements. And due to the sensitivity of the measurement apparatus, the thermal motion of the atoms in the resonant bar gives rise to spikes of comparable magnitude, potentially leading to false positive observations. Similarly, when scientists study human behavior, large sample sizes are needed to average out random variation, and to ascertain the presence or absence of a behavioral effect (Marek et al. 2022). A third common source of discordance is *systematic bias* in observation, measurement, and interpretation: due to methodological choices in sampling, measurement, and data processing, one or more of the conflicting pieces of evidence may systematically misrepresent the property being studied. Typical sources of bias include incorrectly calibrated measurement instruments and unreflectively applied statistical techniques.

This threefold classification of the sources of discordance is not exhaustive, but it provides a starting point for disentangling the complexities involved in the controversies described in the two cases in Section 2.[4] It is likely that all three sources of discordance were present in gravitational wave physics and in priming research. First, the consequences of a non-systematic error by Weber compromised his reputation among his physicist colleagues. In an attempt to rule out alternative explanations of his findings, Weber ran a coincidence experiment where he correlated the time series from his own resonant bar with another one at a different university. The experiment resulted in strongly positive findings, but later it turned out that, due to different time zone settings, the measurements Weber took to be simultaneous were actually recorded four hours apart. Weber had found signal in data where there was no signal to be found (Collins 2004, Ch. 9).

A major point of contention in the gravitational wave controversy centered on the signal detection process: the analytical techniques Weber utilized to discern spikes caused by gravitational waves from random variation, or *noise*, were criticized by other research groups. In general, especially when sample sizes are small, statistical estimates of parameters of interest always involve some sampling error, and there is a non-zero probability of mistakenly detecting a phenomenon in what is actually random variation. The problem is well recognized in psychology, where insufficiently powered methods and small sample sizes, combined with a problematic utilization of "researcher degrees of freedom," have been

[4] As mentioned above, data can function as evidence for a hypothesis only when combined with background assumptions. Therefore, the discordance between the evidence produced by two lines of inquiry can also be attributed to different sets of background assumptions. This viewpoint is compatible with the classification in terms of causes of discordance presented here. We return to this issue in Section 5 where we discuss the relationship between lines of evidence and auxiliary assumptions.

suggested as an explanation of why several findings in experimental psychology have failed to replicate (Simmons et al. 2011; see Meehl 1990, p.125).

These two factors do not explain all of the discordance in the examples. In both controversies, there remains a disagreement between parties who consistently, over repeated studies, claim to reach reliable conclusions that are inconsistent with the findings of the other group. Despite criticisms, and after admitting his errors, Weber remained steadfast in his belief that he had observed gravitational waves, that his findings could not be explained away as results of errors or random variation – much like Bargh (2012a, b). This part of the evidential conflict is what we refer to as persistent discordance.

Collins (1985) argues that scientific controversies are typically not resolved by a rational "algorithm," but by negotiation. Instead of one party being definitively proven wrong, over time both the researchers and the audience simply shift their attention to other questions. In other words, the social controversy may be resolved while the evidential discordance persists (cf. Beatty 2006). Neither non-systematic error nor noise appear as likely explanations of such non-transient discordance, where the epistemic conflict may remain even when the social controversy disappears. In our examples, all of the discrepancies between findings from different studies did not disappear simply by gathering a larger dataset, or by correcting a non-systematic error in the data collection process. To put it crudely, even if two disagreeing groups kept collecting more data, one's findings would still support *H* and the other's not-*H*.[5]

We suggest that the most plausible explanation for persistent discordance is neither non-systematic error or noise but bias, understood in a broad sense. In analogy with the notions of bias and variance in statistics (Wasserman 2004) we use the term "bias" to refer to a systematic discrepancy between the outcome of a study, and the true state of the studied property.[6] Unlike in the case of non-systematic error and noise, averaging several biased trials or studies will not correct the error. Given two series of experiments which result in conflicting findings, typical sources of persistent discordance include the biased functioning of measurement instruments or data analysis procedures. The following sections give several examples of sources of bias in the evidence generation process in both gravitational waves research and social priming studies.

[5] A persistent discordance need not be permanent. The gravitational wave case is an example of a long-term discordance that, despite several steps toward resolution in the 70s (see Sections 4 and 5), was finally resolved in 2015 by the introduction of new kind of evidence from the LIGO experiment.

[6] See Section 4 (the principle of uniformity of nature) and Section 5 (identification of queries and phenomena) for a discussion of the metaphysical commitments underlying our view of evidence generation.

4. Diagnosing Discordance

We argue that the kind of discordance generated by bias cannot be mitigated with the strategies applicable to non-systematic-error- and noise-generated discordance. Furthermore, several philosophical accounts of evidential reliability struggle with discordance generated by bias (see Woodward 2000). According to Woodward, data should be considered reliable when they are counterfactually sensitive to variations in the phenomenon. It is, however, entirely possible for two experimental procedures, leading to discordant outcomes, to manifest such counterfactual sensitivity. Thus, by Woodward's criterion, both would be reliable, but the body of evidence generated by them would be discordant. We argue that uncovering and explicating the discrepancies between two or more *lines of evidence* (Boyd 2018) or chains of *inscriptions* (Latour 1999) embodied by the conflicting research processes can shed light on and help diagnose such cases of persistent discordance.

When faced with evidential discordance, a rational way to proceed often involves temporarily suspending judgment and, if possible, collecting a larger sample of data. Persistent discordance due to bias is problematic because, as argued above, it cannot be resolved simply by suspending judgment and waiting for more data to come in. Unlike discordance due to non-systematic error and noise, the evidential divergence underlying persistent discordance might appear to breach a version of a fundamental assumption science is based on, the principle of uniformity of nature (see Hume 1978, 1.3.6.4). The principle is a presupposition of the extant discussions on discordant evidence and replication. Here we formulate it as follows: if the same query Q is posed to reality several times under similar conditions, nature should, in principle, provide the same answer. In a slightly weaker stochastic formulation, the answers given by nature should be drawn from the same distribution. Our formulation is meant to be metaphysically light, and reflect only a commitment to the idea that there exist stable, identifiable properties in the world, the measurement or determination of which is independent of a single inquirer. In other words, scientists encounter common, intersubjectively shared phenomena. If this were not the case, much of science as well as previous discussions of uncoordinated measurement, replication, and theoretical unification would lose meaning. Likewise, the notion of evidential discordance presupposes the principle: if not true, divergent results from measurements under the same conditions would not be an epistemic problem. The findings would not be discordant but simply different.[7]

How should we then deal with the situation where nature appears to systematically give different answers to a query Q under similar conditions, i.e. persistent discordance? It seems that there are two possible explanations: either there is an error in how one (or more) of the researchers (or research groups) generates an answer to Q, or the queries formulated by the different researchers were not the

[7] We thank a reviewer of this article for pressing us to clarify the relationship between discordance and PoU.

same ones to begin with (cf. Tal 2019). In other words, a study may provide an erroneous answer to a query, or it may misidentify the query it provides an answer to. We address the first possibility below and return to the second in Section 5.

As tentatively suggested in Section 3, discrepancies in study outcomes between different researchers could be attributed to variations in data collection, processing, and interpretation – all processes that reflect the set of background assumptions that a researcher holds. Together, such differences give rise to a discrepancy we call bias. Hence, to rationally address a case of persistent discordance one needs information about the details of that process: how we get from raw data (cf. Gitelman 2013), material items such as recorded specimens or readings from a measuring instrument, to evidence, a propositional evidential claim that stands in a confirmatory relationship with a particular answer to the query *Q*. In the rest of this article, we argue that information regarding this data pipeline (i.e., the evidence generating process) allows us to answer the *descriptive* question of how different researchers or research groups were led to different evidential claims. Furthermore, we argue that to answer the *normative* question about what data pipeline(s) are adequate for providing an answer to a particular query, we must refer to the notion of metadata, i.e., information about how evidence was generated.

Nora Boyd (2018) captures both these aspects in the concept of *enriched line of evidence*: the processes of data generation and transformation in a line of inquiry, enriched with second-order data about said processes. We refer to this second-order data as metadata, defined by Boyd as information "regarding the provenance of the data records and the processing workflow that transforms them" (p. 407). For Boyd, a line of evidence means "a sequence of empirical results including the records of data collection and all subsequent products of data processing generated on the way to some final empirical constraint" (p. 408). The notion of a line of evidence captures the idea that evidence generation is a material, step-by-step process, which can be thought of as a metaphorical line that begins with the research question and ends in an output, a piece of evidence that stands in some relation to a hypothesis, theory, or claim. Together, a line of evidence and the auxiliary information (second-order data, metadata) about the process of evidence generation form an enriched line of evidence.

Although his work comes from a different tradition, we claim that the notion of line of evidence is complemented by Bruno Latour's work on primary and secondary inscriptions (cf. Lewis and Bartlett 2013). Latour (1999) describes how soil scientists on a Brazilian savanna create and transform data about their phenomenon of interest into evidence for a hypothesis. In our vocabulary, the *query* addressed by the research team concerns whether at a particular site, at the boundary between the Boa Vista forest and savanna, it is the forest or the savanna that is advancing into the other. To answer the query, the scientists study the soil at the border between the two. Latour shows how the investigation takes the form of a step-by-step process where material stuff, soil, is transformed into (still material)

data about the soil, which are *turned into evidence* about the movement of the boundary between the two ecological regions. Material soil samples systematically arranged in a research instrument, the "pedocomparator," are data, not evidence (see Section 3). They are transformed through a sequence of steps where clumps of soil and pieces of vegetation are transported from field to office, systematically arranged, classified by comparing them to calibrated color samples, and ultimately turned into numbers and sentences on sheets of paper. The numbers, when finally assembled into a graph, are used as evidence for a particular answer to the query *Q*. Such a sequence of causal transformations can be thought to constitute the line of evidence leading from data to an evidential claim.

Thinking about scientific evidence in terms of lines suggests a solution to the puzzle of why nature sometimes appears to give different answers to the same question. If two scientists study the same lump of soil with the same research question in mind, but nevertheless repeatedly end up with contradictory evidence, the discrepancy may be due to the differences between the processes with which some property of the phenomenon is captured in data, and the data used as evidence for a particular answer to *Q*. Here too, discordance is a result of systematic bias embodied in one or both of the lines of evidence. Notice how viewing the depth of the process between the initial measurement and the ultimate evidential claim provides a more applicable approach to understanding discordance than a traditional "flat" view of confirmation involving just the hypothesis and the evidential claim, or as conceptualizing discordance as an instance of underdetermination.

The contradictory conclusions reached by Weber and his critics can at least partly be attributed to the differences between the lines of evidence created by the experimental setups. Potentially significant differences in design can be found at various stages of the line of evidence. First, the groups used *different means of detecting vibration in the resonant bar*: Weber used piezo-electric crystals glued to the bar. Other labs used capacitors whose plate separation changed with changes in the length of the bar. The second factor is the *material of the bar*: most experiments used aluminum alloy, but some others pure aluminum or huge crystals of sapphire. Third, different teams used *different electronics for signal processing*. Fourth, *different analytical techniques* were used to extract signal from noise. For detecting peaks, Weber's group relied on a non-linear algorithm that was sensitive only to the amplitude of the signal, whereas other groups employed a linear algorithm that made use of changes in both amplitude and phase (Franklin 1997). Each of the four differences merits critical attention, as they may provide a clue to the source of persistent evidential discordance.[8]

[8] Based on historical case studies, Hudson (2014) argues that scientists engage in "reliable process reasoning" (not mere robustness reasoning) in forming scientific opinion. Our approach suggests that careful scrutiny of the different stages of a line of evidence provides an epistemic strategy for such reliability assessment.

There is less openly available information about the lines of evidence in the social priming controversy, but it too can be approached by attempting to localize the source of the discordance in the lines of evidence. For example, although in their replication attempt Doyen and colleagues tried to match the original experimental setup, they did introduce some changes. They doubled the number of participants in the experiment, timed their walking speed with infrared sensors rather than a stopwatch, and enhanced experimental blinding by recruiting experimenters without informing them about the aim of the experiment. Each of these changes can be seen as an attempt to look "under the hood" of Bargh's evidence generation process and eliminate sources of error in the line of evidence: a larger sample size decreases noise, and automatized timing and appropriate blinding can eliminate systematic bias resulting from experimenter expectations.

5. Resolving Persistent Evidential Discordance

Uncovering lines of evidence provides an approach to locating the source(s) of discordance between prima facie similar setups that generate contradictory findings. We still need to address the normative question of correctness: Given that the conclusions from different studies indeed diverge, how do we find out which study provides a legitimate answer to the original query? We contend that escaping the epistemic impasse exemplified by the experimenter's regress requires an answer to the question of how causal differences between lines of evidence get turned into normative differences in evidential warrant.

We believe that the metadata included in an enriched line of evidence is a crucial resource here, as it specifies what each transformation step in the line of evidence *is supposed to do.*[9] While many fields like computer science and data science employ the term 'metadata' in a narrow technical sense, we adopt a broader definition, understanding it simply as data about data, devoid of particular technical constraints related to its formatting or representation. Like Boyd (2018, p. 411) we find it useful to think of enriched lines of evidence in analogy with the notion of an ideal explanatory text (Railton 1981). Although in practice metadata is often scattered and full of gaps, it ideally contains descriptions of each of the transformations implemented in the line of evidence. In his case study Latour (1999) found that the soil scientists both consume and generate metadata: they collect and transform their samples carefully following a pre-specified protocol, and while doing so, they record their activities in logbooks. Carefully curated lab notebooks have long been recognized as an important metadata practice guaranteeing the reliability of research. Similarly, the operating principles of new online platforms for reproducible research and data sharing also reflect increasing attention to good metadata practices.

Note that without metadata, data alone cannot be used as evidence for anything; a spreadsheet filled with numerical results from a series of measurements but no information about the instruments used,

[9] See Bokulich (2020a) for a classification of data transformation steps.

time of measurement or location could not be used as evidence for a scientific claim (see Wylie 2020, p.294). A carefully curated and well documented dataset, in contrast, can sometimes be detached from its original evidential context and be used as evidence to address new queries (cf. Leonelli and Tempini 2020).

Metadata, when understood as an ideal text as described above, provides the normative benchmark for each causal transformation step in the line of evidence. For example, if a measuring instrument is chosen because of its capacity to detect a signal with error smaller than ε, and the actual error turns out to be an order of magnitude larger, this surely compromises the correctness of the processing conducted in that step. Similarly, if a numerical method used by a statistical software package does not reliably converge to the correct value, it fails to meet its normative standard, and hence the reliability of the processing step is called into question. This is one way in how referring to metadata allows one to move from the descriptive to the normative domain, and to speak of the (in)adequacy of a line of evidence for addressing a particular query.[10]

To further illustrate this juxtaposition of the causal and the inferential, we can think of the normative standard for a line of evidence as consisting of a sequence of mappings $f_i: X_i \rightarrow X_{i+1}$, each representing a data transformation step that maps elements of stage i (set X_i) in the line of evidence to elements of X_{i+1}. For example, Latour (1999) describes how the soil studied by pedologists (stage i) gets transformed into a set of soil samples systematically organized in a suitcase-like storage instrument, the pedocomparator (stage $i+1$). By using a compass, angle measurements, and a simple instrument called a topofil for measuring distances in the wild, the researchers superimpose a grid of Cartesian coordinates on the landscape. This allows them to systematically drill soil samples, label each sample with coordinates, and to insert it in the pedocomparator grid. This rule-governed processing of data can be seen as a material instantiation of a particular mapping f_i. Later on, in the laboratory, the collection of samples in the pedocomparator organized in rows and columns – a representation of soil in the region studied – is used as the raw material for further processing steps.

Generally, in such a sequence of transformations, elements from one stage, such as observations, are mapped onto the next, such as a quantitative representation of those observations.[11] Each step typically involves filtering, data compression with loss, but in a way that captures some inferentially significant invariance (represented by f_i) in the data. However, if the actual causal transformation in stage i cannot

[10] We believe that our notion of metadata corresponds to what Tal (2019) calls a model of the measurement process. Both our notion and Tal's model function as a precondition for getting objective knowledge from a causal process, and allow scientists to differentiate genuine results from artifacts.

[11] Following the notation given above, the normative benchmark for the entire line of evidence can be represented as a composite function: $\mathcal{F} = f_n \circ \ldots \circ f_2 \circ f_1$

be seen to instantiate f_i, i.e., there is a discrepancy between the metadata and the causal process of data transformation, the evidential status of data produced by such a step is undermined. Because later stages of a line of evidence take outcomes of earlier ones as input, often even a single non-negligible discrepancy compromises the reliability of the answer given to Q.

Now let's turn to another way in which attention to metadata allows comparing competing lines of evidence. Consider fitting a linear model to a set of observations. It too can be seen as one step along the line of evidence. A linear model enables one to describe a set of data points – of any size, the more the better – with just a few parameters. Such data compression does not obviously come for free: the linear model can only capture an invariance in the data if the data is, indeed, linear. Hence, only when combined with the linearity assumption can the parameters of the fitted model later be used as reliable evidence of the relationship between the measured variables. In this sense, each additional mapping f_i could be seen to require additional auxiliary assumptions – often innocent but sometimes substantial – the conjunction of which is needed in the argument that uses the final stage of the line of evidence to answer the query Q.[12]

In practice, all complex experimental setups must rely on a host of background assumptions that justify the design of the line of evidence. For example, Weber's experimental setup relied on assumptions concerning the length of the bursts of radiation (the detectors could not detect very short bursts) and proximity to the source of radiation. The frequency distribution of the radiation was another crucial assumption: the bar in Weber's setup could only be built to be sensitive to a narrow frequency range. Consequently, the setup depended on the assumption that the band around 1660Hz was a plausible range for detecting gravitational waves. At the time, no direct evidence concerning the plausibility of such assumptions was available, and therefore, there was no direct way to evaluate the adequacy of the line of evidence in Weber's experiment. Arguments against Weber were indirect: his findings about how often gravitational waves were observed were shown to imply an implausibly high rate of energy consumption. If Weber's findings were correct, the universe would use up its energy much faster than suggested by other means of estimation.

In the social priming controversy, the plausibility of the assumptions underlying Doyen's experimental design was openly contested by Bargh (2012a). According to him, Doyen's changes to the setup made it impossible to observe the priming effect.[13] In general, the argument from false auxiliaries is easier to

[12] For a visual illustration of auxiliary assumptions being combined with the evidential claim, see Figure I.2 in Cartwright and Hardie (2012).

[13] In so doing, Bargh applies the Lakatosian negative heuristic, where a falsification attempt is deflected to a false auxiliary assumption (see Meehl 1990). In order to avoid turning into mere ad hoc moves characteristic of a degenerating research program, such defenses of – in this case – a line of evidence should be counterbalanced by some dimensions of evidential value (a solid track record of successful empirical use, "money in the bank,"

make against failed replications than successful ones. Not detecting a phenomenon is almost always easier than detecting it: whereas successful detection requires all of the relevant auxiliaries to be true, only a single false auxiliary assumption can be sufficient to obtain a false negative result (cf. Strack 2017). To respond to such an argument, the author of a failed replication needs to convincingly argue for the sensitivity of their procedure. That is, if the phenomenon existed, the procedure would detect it (see Woodward 2000; Franklin 1997).

The examples above suggest that the steps toward resolving persistent discordance may build on epistemic evaluation of lines of evidence in two ways. First, it may turn out that the line of evidence does not function as specified by its normative benchmark (metadata). Second, the design of the line of evidence may turn out to reflect background assumptions that compromise its capacity to provide reliable answers to the query *Q*.[14,15] The historical development of the social priming controversy suggests yet another direction that the resolution of persistent discordance may take. It is not uncommon for the original query to be somewhat ambiguous, and hence difficult to connect to a specific set of auxiliary assumptions embodied in an experimental setup. For example, recall from Section 2 that priming is generally characterized as the phenomenon of environmental, observational, and perceptual stimuli pre- or non-consciously activating behavioral outcomes. It is far from obvious how different queries concerning such a phenomenon should be identified.

This draws attention to the care needed in discussions of experimental replication. An exact replication is elusive, experimental conditions can never be exactly repeated or reproduced (Popper 2005/1959, 420–422; Machery 2020; Feest 2019). The experimental setup used by Bargh's group on the one hand, and the alternative one utilized in the failed replication on the other, can be seen to embody different sets of experimental assumptions which, in fact, address slightly different queries. For example, we can imagine Bargh granting that Doyen's experiments were run correctly, i.e., according to the process described by the ideal metadata of the process, and, indeed, that no priming would be detected in such experiments. Bargh could then continue on to argue that such an experiment should be seen as not answering the original query *Q* but a different one, *Q**, which asks whether priming effects would be detected under conditions specified by Doyen and colleagues. Bargh could sideline *Q**, and answers given to it, not as incorrect, but as *irrelevant* to his original query *Q.* Discordance morphs into a debate

and potential to contribute new answers to queries of interest). We thank a reviewer of this article for pointing out the connection to Meehl's work.

[14] Stegenga (2018, Ch 7) gives examples of evidential discordance in the context of medicine. Stegenga's examples could be understood as involving evaluation of lines of evidence under considerable uncertainty.

[15] Attention to the background assumptions embodied in competing lines of evidence makes it possible to compare them according to several dimensions criteria of evidential value. For example, whereas one line of evidence may generate large amounts of noisy data, another could provide more precise data (see Bokulich & Parker 2021), but at a slower speed.

concerning relevance (of queries and answers given to them).[16] The question becomes: Which line provides an answer to the original query *Q* – or was *Q* well-formulated in the first place? Which line of evidence is more relevant to what we ultimately want to know?

In the social and behavioral sciences, one strategy to address such metalevel controversy is to take a modest stance on the external validity of findings. Instead of searching for universal patterns in human behavior, advocates of the heterogeneity revolution emphasize the unavoidable locality of findings in the behavioral sciences (Bryan, Tipton and Yeager 2021). The findings from psychological research may be less generalizable and more sensitive to moderating variables than previously thought, but this should be understood rather as epistemic humility than as an argument undermining the relevance of carefully conducted research.[17] In such situations, persistent discordance may be resolved as the original, general but loosely formulated, query gets replaced with several more local but more carefully specified ones, each answered to by a different line of evidence.

6. Conclusions

We have argued that uncovering information about enriched lines of evidence provides a strategy for addressing the problem of persistent evidential discordance. By exposing the process used to generate an evidential claim, it is possible to break the experimenter's regress and to locate the source of persistent discordance. The metadata included in the enriched line of evidence provide the normative standard against which the causal processing steps in the line of evidence can be evaluated.

Discordance due to noise can readily be accounted for by understanding the properties of the lines of evidence in question. Likewise, discordance caused both by non-systematic error and systematic malfunctioning of (the hardware and software) machinery can be explained as clear departures from how data are supposed to be processed, as specified by metadata. Persistent discordance may, however, also arise when two competing lines of evidence function as specified by their normative inferential standard, and yet produce contradicting results. In such situations, the disagreement turns into one of evidential relevance and the epistemic virtues of different lines of evidence, which must be dealt with by evaluating the plausibility of the auxiliary assumptions embodied in each line with regard to the query *Q*.

[16] Not only have epistemic values but also social values been argued to influence judgments of evidential relevance (Longino 2022). In such a view, values may act as background assumptions influencing, for example, the evaluation of the relevance of queries.

[17] See Ramscar (2016) on the population-sensitive nature of social priming phenomena and its methodological implications.

In most cases, uncovering the enriched line of evidence in its entirety is not a tractable goal. Having a line of evidence fully in sight should be considered an ideal situation, guiding rational diagnosis of evidential discordance. Finally, although discordance is a problem for scientific consensus formation and truth, as well as a serious roadblock for those who wish to use science for decision-making, discordance can play a significant role in the broader dynamics of inquiry. Discordance can only occur when a scientific research question has been addressed by more than one study – when a replication has been attempted (intentionally or unintentionally). The failed replication calls for explanation and hence discordance acts as "convenient friction," drawing attention and resources to the scientific issue at hand (Price 2003; cf. Ohnesorge 2021, Bokulich 2020b). The resolution of discordance – whether it leads to more accurate findings or more nuanced queries – is a necessary step in the self-corrective process of scientific research.

Acknowledgments

The authors wish to thank the two anonymous referees for their insightful and constructive feedback, and Kate Sotejeff-Wilson for language editing. The authors also wish to thank audiences at PSA22, the European PhD Network for Theoretical Philosophy and Philosophy of Science workshop, and the Brown Bag and Perspectives on Science research seminars at TINT for their useful feedback. Samuli Reijula's research is funded by the Research Council of Finland (#332686). Blanco Sequeiros' research is funded by the Kone Foundation (#20190173).

Samuli Reijula
Theoretical Philosophy / TINT
University of Helsinki
Helsinki, Finland
samuli.reijula@helsinki.fi

Sofia Blanco Sequeiros
Practical Philosophy / TINT
University of Helsinki
Helsinki, Finland
sofia.blancosequeiros@helsinki.fi

References

Abbott, B. P., Abbott, R., Abbott, T. D., Abernathy, M. R., Acernese, F., Ackley, K.,… & Cavalieri, R. (2016). "Observation of gravitational waves from a binary black hole merger." *Physical review letters*, 116(6), 061102.

Achinstein, P., (2001). *The book of evidence.* Oxford University Press.

Bargh, J.A. (2012a). "Nothing in their heads. Debunking the Doyen et al. claims regarding the elderly-priming study." URL: https://replicationindex.com/wp-content/uploads/2020/07/bargh-nothingintheirheads.pdf

Bargh, J. A. (2012b). "Priming effects replicate just fine, thanks." *Psychology Today*, *11*.

Bargh, J.A., Chen, M. & Burrows, L., (1996). "Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action." *Journal of Personality and Social Psychology*, 71(2): 230–244.

Bargh, J.A. & Chartrand, T.L., (1999). "The unbearable automaticity of being." *American psychologist*, 54(7): 462–479.

Beatty, J. (2006). "Masking disagreement among experts." *Episteme*, *3*(1–2), 52–67.

Bogen, J. & Woodward, J. 2005. "Evading the IRS." In M. R. Jones & N.Cartwright (eds.), *Idealization XII: Correcting the Model: Idealization and abstraction in the sciences.* Rodopi.

Bokulich, A. (2020a). "Towards a taxonomy of the model-ladenness of data." *Philosophy of Science*, 87(5), 793–806.

Bokulich, A. (2020b). "Calibration, Coherence, and Consilience in Radiometric Measures of Geologic Time." *Philosophy of Science*, 87(3), 425–456.

Bokulich, A., & Parker, W. (2021). "Data models, representation and adequacy-for-purpose." *European Journal for Philosophy of Science*, 11, 1–26.

Bovens, L., & Hartmann, S. (2004). *Bayesian epistemology*. OUP Oxford.

Boyd, N.M., (2018). "Evidence enriched." *Philosophy of Science*, 85(3): 403–421.

Breznau, N., Rinke, E. M., Wuttke, A., Nguyen, H. H., Adem, M., Adriaans, J.,... & Van Assche, J. (2022). "Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty." *Proceedings of the National Academy of Sciences*, 119(44), e2203150119.

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). "Behavioural science is unlikely to change the world without a heterogeneity revolution." *Nature Human Behaviour*, *5*(8), 980–989.

Cartwright, N. (2013). "Evidence, argument and prediction." In *EPSA11 perspectives and foundational problems in philosophy of science* (pp. 3–17). Springer, Cham.

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

Cesario, J. (2014). "Priming, replication, and the hardest science." *Perspectives on psychological science*, 9(1), 40–48.

Chapman, R., & Wylie, A. (2018). *Evidential reasoning in archaeology.* Bloomsbury Publishing.

Claveau, F., (2013). "The independence condition in the variety-of-evidence thesis." *Philosophy of Science*, 80(1), 94–118.

Chivers, T. (2019). "What's next for psychology's embattled field of social priming." *Nature,* 576(7786): 200–203.

Collins, H., (1985). *Changing order: Replication and induction in scientific practice.* University of Chicago Press.

Collins, H. (2004). *Gravity's shadow*. University of Chicago Press.

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). "Behavioral priming: it's all in the mind, but whose mind?". *PloS one*, *7*(1), e29081.

Earp, B. D., & Trafimow, D. (2015). "Replication, falsification, and the crisis of confidence in social psychology." *Frontiers in psychology*, 6, 621.

Feest, U. (2019). "Why replication is overrated." *Philosophy of Science*, 86(5), 895–905.

Franklin, A. (2002). *Selectivity and discord: Two problems of experiment.* University of Pittsburgh Press.

Franklin, A. (1997). "Calibration." *Perspectives on Science*, *5*(1), 31–80.

Gitelman, L. (Ed.). (2013). *Raw data is an oxymoron*. MIT press.

Harris, C. R., Coburn, N., Rohrer, D., & Pashler, H. (2013). "Two failures to replicate high-performance-goal priming effects." *PloS one*, *8*(8), e72467.

Hey, S.P., (2015). "Robust and discordant evidence: Methodological lessons from clinical research." *Philosophy of Science*, 82(1): 55–75.

Hudson, R. (2014). *Seeing things: The philosophy of reliable observation*. Oxford University Press, USA.

Hume, D. (1978). *Treatise of human nature* (L. A. Selby-Bigge, Ed.; 2nd ed.). Oxford University Press.

Kelly, T. (2016). "Evidence", *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), E. N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/evidence/>.

Kuorikoski, J., & Marchionni, C. (2023). "Evidential Variety and Mixed-Methods Research in Social Science." *Philosophy of Science*, 90(5), 1449-1458.

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

Latour, B., (1999). "Circulating reference." In *Pandora's hope: Essays on the reality of science studies*, 24–79. Harvard University Press.

Leonelli, S. (2019). *Data-centric biology: A philosophical study.* University of Chicago Press.

Leonelli, S., & Tempini, N. (eds.) (2020). *Data journeys in the sciences*. Springer Nature.

Lewis, J. & Bartlett, A., (2013). "Inscribing a discipline: Tensions in the field of bioinformatics." *New Genetics and Society*, 32(3): 243–263.

Lin, H. (2022). "Bayesian epistemology", *The Stanford Encyclopedia of Philosophy* (Fall 2022 Edition), E.. Zalta & U. Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2022/entries/epistemology-bayesian/>.

Longino, H. E. (2022). What's Social About Social Epistemology?. *The Journal of Philosophy*, 119(4), 169-195.

Machery, E. (2020). What is a replication? *Philosophy of Science*, 87(4), 545–567.

Marek, S., et al., (2022). "Reproducible brain-wide association studies require thousands of individuals." *Nature* 603: 654–660.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. University of Chicago Press.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological inquiry*, 1(2), 108–141.

Ohnesorge, M. (2022). "Pluralizing measurement: Physical geodesy's measurement problem and its resolution." *Studies in History and Philosophy of Science A*, 96, 51–67.

Ohnesorge, M. (2021). "How incoherent measurement succeeds: Coordination and success in the measurement of the earth's polar flattening." *Studies in History and Philosophy of Science Part A*, *88*, 245–262.

Oreskes, N., & Conway, E. M. (2011). *Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming*. Bloomsbury Publishing USA.

Peirce, C. S., (1878) [1986], "How to Make Our Ideas Clear", *Popular Science Monthly*, 12(January): 286–302; reprinted in *Writings of Charles S. Peirce: A Chronological Edition* (Volume 3: 1872–1878), C. Kloesel, M. Fisch, N. et al. (eds.), Bloomington, IN: Indiana University Press, 257–276.

Popper, K. (2005/1959). *The logic of scientific discovery*. Routledge.

Price, H. (2003). "Truth as convenient friction." *The Journal of Philosophy*, *100*(4), 167–190.

Railton, P. (1981). "Probability, Explanation, and Information." *Synthese* 48 (2): 233–56.

Ramscar, M. (2016). Learning and the replicability of priming effects. *Current Opinion in Psychology* 12, 80-84.

Romero, F. (2019). "Philosophy of science and the replicability crisis." *Philosophy Compass*, 14(11), e12633.

Shwed, U., & Bearman, P. S. (2010). "The temporal structure of scientific consensus formation." *American sociological review*, 75(6), 817–840.

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.

Smith, G. E. (2014). "Closing the loop." In Biener and Schliesser (Eds.) *Newton and empiricism*, pp. 262–352.

Stegenga, J. (2018). *Medical nihilism*. Oxford University Press.

Stegenga, J., (2012). "Rerum concordia discors: Robustness and discordant multimodal evidence." In *Characterizing the robustness of science*, 207–226. Springer, Dordrecht.

Stegenga, J., (2009). "Robustness, discordance, and relevance." *Philosophy of Science*, 76(5): 650–661.

Strack, F. (2017). "From Data to Truth in Psychological Science. A Personal Perspective." *Frontiers in psychology*, 702.

Tal, E. (2019). "Individuating quantities." *Philosophical Studies*, 176(4), 853–878.

Wasserman, L. (2004). *All of statistics: A concise course in statistical inference* (Vol. 26). New York: Springer.

Weingarten, E., Chen, Q., McAdams, M., Yi, J., Hepler, J., & Albarracín, D. (2016). "From primed concepts to action: A meta-analysis of the behavioral effects of incidentally presented words." *Psychological bulletin*, 142(5): 472–497.

Woodward, J. (2000). "Data, phenomena, and reliability." *Philosophy of Science*, *67*, S163-S179.

Wylie, A. (2020). Radiocarbon Dating in Archaeology: Triangulation and Traceability. In: Leonelli, S., Tempini, N. (eds.) *Data journeys in the sciences*. Springer, Cham.

Yong, E., (2012). "Nobel laureate challenges psychologists to clean up their act." *Nature,* 490: 7418.