Chapter 9
Rerum Concordia Discors: Robustness and Discordant Multimodal Evidence

Jacob Stegenga

But to stand in the midst of this rerum concordia discors and of this whole marvelous uncertainty and rich ambiguity of existence. . .
Nietzsche, Gay Science I.2

A symphony of Beethoven presents to us the greatest confusion, which yet has the most perfect order at its foundation, the most vehement conflict, which is transformed the next moment into the most beautiful concord. It is rerum concordia discors, a true and perfect picture of the nature of the world which rolls on in the boundless maze of innumerable forms. . .
Schopenhauer, Metaphysics of Music

Quid velit et possit rerum concordia discors. Empedocles deliret acumen?
What does the discordant harmony of things mean, and what can it do? Is Empedocles crazy?
Horace, Epistles I.12.19

9.1 Introduction: Multimodal Evidence

We learn about particular aspects of the world with multiple methods. Galileo's defense of heliocentrism was based on late-sixteenth century astronomical novelties, Brahe's naked-eye observations of Mars and Kepler's accounting of them with elliptical orbits, and Galileo's own telescopic observations of Jupiter's moons and shifting sunspots. Evidence mustered to support Wegener's theory of continental drift included paleontological parallels between continents, stratigraphic parallels between continents, and the jigsaw-puzzle fit of continents. When Avery and his colleagues suggested that genes might be composed of deoxyribonucleic

J. Stegenga (✉)
University of Toronto, Toronto, ON, Canada
e-mail: jacob.stegenga@utoronto.ca

L. Soler et al. (eds.), *Characterizing the Robustness of Science*, Boston Studies in the Philosophy of Science 292, DOI 10.1007/978-94-007-2759-5_9,
© Springer Science+Business Media B.V. 2012

acid (DNA), their evidence included chemical analysis, enzymatic experiments, ultraviolet absorption, electrophoresis, and molecular weight measurements.[1] When Tom Ridge was the governor of Pennsylvania he signed the death warrant of Mumia Abu-Jamal, who is accused of shooting a police officer and now sits on death row; Abu-Jamal's purported guilt is supported by testimony of four direct witnesses, and the retrieval of his gun and spent cartridges at the murder scene, which matched the bullets extracted from the murdered officer.

Galileo also had to consider contrary evidence: that bodies fall straight to earth, for example, and evidence of an altogether different kind – the authority of sacred texts – since Ecclesiastes tells us that "the sun also rises." There was evidence against continental drift: for example, data indicated that the Earth's mantle is rigid. Evidence that proteins are the functional basis of genes, rather than DNA, was also manifold: proteins are sufficiently diverse in structure and function to be the basis of heredity, in contrast to the supposed simplicity of DNA in the 1940s, and it was highly probable that Avery's samples of DNA also included undetected contaminating protein. The purported innocence of Abu-Jamal was supported by testimony from multiple witnesses, and after the original guilty verdict (but before Ridge's condemnation) the admission of someone else as the killer, and other information that suggested that much of the original prosecution evidence was flawed.

Galileo, Wegener, Avery, and Ridge relied on "multimodal evidence." Some call this "evidential diversity" (e.g. Fitelson 1996). That term is fine with me. It is a noun. Mine is an adjective – multimodal – and my modified noun is "multimodal evidence". This is a useful neologism because it allows talk of individual modes of evidence and their various relations to each other and to different hypotheses. It is also a salient neologism because it calls to mind our sensory modalities; much sensation is literally multimodal evidence. Locke argued that we are more likely to believe in primary qualities because we observe them with multiple sensory modalities, as when we observe extension with both touch and sight, whereas secondary qualities, like color, we observe only with a single sensory modality. What I mean by "mode" is a particular way of finding out about the world; a type of evidence; a technique or a study design. The total set of evidence that is relevant to a hypothesis of interest and that is generated by multiple modes I call *multimodal evidence*.

When multimodal evidence for a hypothesis is concordant, that is often said to be epistemically valuable.[2] Evidence that is varied is said to provide more support to a hypothesis than does homogeneous evidence. This is how Hempel put it: "The confirmation of a hypothesis depends not only on the quantity of the favorable evidence available, but also on its variety: the greater the variety, the stronger the

[1] See Westman (2011); Oreskes (1999); Stegenga (2011).

[2] Many philosophers of science have claimed that concordant multimodal evidence is useful, including Hempel (1966), Wimsatt (1981), Horwich (1982), Cartwright (1983), Hacking (1983), Franklin and Howson (1984), Howson and Urbach (1989), Trout (1993), Mayo (1996), Achinstein (2001), Staley (2004), Chang (2004), Douglas (2004), Allamel-Raffin (2005), Weber (2005), Bechtel (2006), Kosso (2006). The contributions to this volume are some of the first to critically evaluate such arguments.

resulting support" (1966, p. 34). Conversely, when multimodal evidence is discordant, that is often said to be conducive to uncertainty – there are several responses one hears: the evidence is too messy to know what to believe; or, the 'weight of the evidence' more strongly suggests this hypothesis over that; or, more research is required. Hempel, in the passage above, only mentioned the quantity and variety of "favorable" evidence, but surely confirmation must depend on both favorable and unfavorable evidence. Without a method of systematically assessing and combining multimodal evidence, both views – that concordant multimodal evidence is a Good Thing, and that discordant multimodal evidence is a Bad Thing – are, as I argue in Section 9.6, unsatisfactory.

Multimodal evidence is an exceptionally important notion: it is ubiquitous in science and law; it elicits both certainty and dissent amongst practitioners; and yet it is poorly understood. The above remarks suggest three questions: (1) determining what multimodal evidence *is*; (2) specifying *why* multimodal evidence is valuable; and (3) describing *how* multimodal evidence should be assessed and combined to provide systematic constraint on our belief in a hypothesis. There is little literature addressing the first question; there have been several answers suggested for the second question, one of the most prominent of which is the notion of robustness; and there are several disputed approaches to the third question.

In Section 9.5 I address the first question, and conclude that determining criteria for defining a mode of evidence is a difficult conceptual problem, the solution to which will likely be relative to the way one wishes to use multimodal evidence (this is what I call the *individuation problem* for multimodal evidence). First, though, I discuss two of the prominent answers to the second question: the notion of 'robustness' is one account of how multimodal evidence is said to be valuable (Section 9.2), and 'security' is another (Section 9.3). My explication of multimodal evidence ends in Section 9.6, where I discuss the challenge of assessing and amalgamating multimodal evidence.

9.2 Robustness

One of the primary ways in which multimodal evidence is purported to be valuable is because concordant multimodal evidence is said to be better evidence for a hypothesis, ceteris paribus, than evidence from a single mode; hypotheses supported by concordant independent multimodal evidence are said to be *robust*. Robustness is a recent term that undergirds a common platitude: hypotheses are better supported with plenty of evidence generated by multiple techniques that rely on different background assumptions. A simple example of this was given by Ian Hacking when he argued that if a cellular structure is observed with different types of microscopes, then there is more reason to believe that the structure is real (1983). I have seen the term "robustness" first used as a methodological adage by the statistician George Box in 1953 – a robust statistical analysis is one in which its conclusions are consistent despite changes in underlying analytical assumptions. In philosophy of science I have seen the term first used with respect to models: results consistent across

multiple models (with different background assumptions) are 'robust' and so more likely to be true (Levins 1966; Wimsatt 1981); Levins' infamous quip is that "our truth is the intersection of independent lies." Nearly every philosopher of science interested in evidence has, at least in passing, extolled the virtues of robustness.

Thus, robustness can be a feature of statistical analyses, models, and hypotheses. My concern in this chapter is with empirical hypotheses.

> *Robustness*: A hypothesis is robust if and only if it is supported by concordant multimodal evidence.

Another name that concordant multimodal evidence has gone by is "independent determinations" (see, for example, Wimsatt 1981 and Weber 2005). The common presumption is that robustness is epistemically valuable, since concordant multimodal evidence provides greater confirmational support to a hypothesis than does evidence from a single mode of evidence. My definition above has an element of independence between lines of evidence, or 'determinations', built in, since the notion of multimodal evidence is assumed to have a criterion of individuation for modes of evidence. However, as suggested in Sections 9.2 and 9.5, determining both *how* and *if* modes of evidence are independent is difficult. The value of robustness is often simply assumed or left implicit, but one way to understand robustness is as a no-miracles argument: it would be a miracle if concordant multimodal evidence supported a hypothesis and the hypothesis were not true; we do not accept miracles as compelling explanations; thus, when concordant multimodal evidence supports a hypothesis, we have strong grounds to believe that it is true.

Robustness is often presented as an epistemic virtue that helps us achieve objectivity. Champions of robustness claim that concordant multimodal evidence can demarcate artifacts from real entities, counter the "experimenter's regress," ensure appropriate data selection, and resolve evidential discordance. Consider the worry about artifacts: if a new technique shows x, the observation of x might be due to a systematic error of the technique rather than due to the reality of x. *Response*: if x is observed with concordant multimodal evidence it is extremely unlikely that x is an artifact (Hacking 1983). Consider the "experimenter's regress": good evidence is generated from properly functioning techniques, but properly functioning techniques are just those that give good evidence (Collins 1985). *Response*: this vicious experimental circle is broken if we get the same result from concordant multimodal evidence (Culp 1994). Consider the concern about data selection: scientists use only some of their data, selected in various ways for various reasons, and the rest is ignored – but how do we know that the selection process gives true results? *Response*: vary the selection criteria, and invariant results are more likely to be true (Franklin 2002). Finally, consider discordant data: multiple experimental results do not always agree – which results should we believe? *Response*: simply conduct more experiments until they yield concordant multimodal evidence.

Robustness has been used as an argument for realism. The canonical example is Jean Perrin's arguments for the reality of atoms (described in Nye 1972 and discussed in Cartwright 1983; Salmon 1984; and Mayo 1996). Jean Perrin calculated Avogadro's number consistently, using different kinds of experiments:

Brownian motion, alpha particle decay, X-ray diffraction, blackbody radiation, and electrochemistry, and the common-cause for this consistency is the existence of molecules.

Given the variety of epistemic tasks placed on robustness, and given the frequency with which the notion is appealed to, it has received surprisingly little direct philosophical evaluation; the chapters in this volume are an important contribution towards understanding the value and challenges of robustness. I will discuss several problems with robustness, in an attempt to provide needed constraints on the concept. Robustness is valuable in ideal evidential circumstances, when all available evidence is concordant. One major difficulty for robustness is that in many cases multimodal evidence is not concordant. When multimodal evidence is available for a given hypothesis, the evidence is often discordant; that is, evidence from various modes supports competing hypotheses. The general applicability of robustness is mitigated by the problem of discordant evidence. Moreover, scientists have some methods for assessing and combining multimodal evidence, but without using such methods in a robustness-style argument, such an argument is at best a pump of one's intuitions justifying a vague or qualitative conclusion.

9.2.1 Three Preliminary Challenges

Prior to discussing what I consider to be the 'hard' problems of robustness – discordance and individuation – I discuss three preliminary challenges. First, scientists do not always have multiple modes of evidence with which to make a robustness-style argument; second, knowing that multiple modes are independent is difficult or impossible (as is knowing in what way multiple modes should be independent; I discuss this in Section 9.5); and finally, concordant multimodal evidence will not necessarily give a correct conclusion. None of these problems taken alone completely repudiates the value of robustness. Indeed, it is a (trivially) important methodological strategy which scientists frequently use. However, the value of robustness is mitigated, and its extent of application constrained, upon consideration of these three preliminary challenges.

Generating concordant multimodal evidence is difficult. Scientists need evidence from independent modes to make a robustness claim, but they do not always have multiple independent modes of evidence to study the same subject. New modes are introduced into scientific practice for good reason: they give evidence on a new subject, or on a smaller or larger scale, or in a different context, than do existing modes. Even if multiple modes do exist, it is not always clear that they are independent. Bechtel (2006) argued that since new techniques are often calibrated to existing techniques even when both techniques provide concordant results the techniques might fail to be independent (see also Soler's discussion on 'genetic non-independence' in the introduction). Furthermore, determining what criteria should be used to determine independence between modes is a difficult problem; this is what I call the "individuation problem" for multimodal evidence (Section 9.5). Simply put, the following challenges must be met to make a robustness argument:

one must have independent modes of evidence, one must have a criterion to which one can appeal in order to demarcate modes of evidence, and one must know that the available modes meet this criterion so that we can be confident that the modes are properly independent. Since robustness requires multiple modes of evidence, an incomplete or vague individuation of evidential modes will render robustness an incomplete or vague notion, and hence robustness-style arguments will be vague or inconclusive.

One might think that multiple invalid arguments that reach the same conclusion give no better reason to believe this conclusion than a single invalid argument reaching the same conclusion. Similarly, multiple methodologically mediocre experiments, or multiple epistemically unrelated experiments, or multiple modes of evidence with implausible background assumptions, give no better reason to believe a hypothesis than does a single mode (let alone a single well-performed mode with more plausible background assumptions). A detailed case-study discussed by Nicolas Rasmussen provided an instance of this problem: multiple methods of preparing samples for electron microscopy demonstrated the existence of what is now considered an artifact (1993). Although this case study generated a good amount of controversy – see responses from Culp (1994), G. Hudson (1999), and others – the fact that such evidential diversity was used as an argument for the reality of an artifact mitigates the epistemic value of robustness. The problem demonstrated by Rasmussen can be generalized: concordant multimodal evidence can support an incorrect conclusion.

In short, to make a compelling robustness argument, one needs evidence from multiple modes for the same hypothesis, while ensuring that such modes are sufficiently independent. Scientists are often adept at grappling with these challenges. However, the problem raised by Rasmussen indicates that arguments based on robustness can generate incorrect conclusions. In other words, robustness requires having multiple modes of evidence, knowing that multiple modes of evidence are independent and knowing how they should be independent, and yet remains fallible. Knowing *that* multiple modes of evidence are independent depends on knowing *how* multiple modes of evidence must be independent to be sufficient for a robustness argument. The former obviously depends on the latter. In Section 9.5 I discuss the latter problem: what I call the individuation problem for multimodal evidence.

9.3 Security

It is a familiar platitude that data is only evidence with respect to a hypothesis, and to think that data is relevant to a hypothesis we must accept certain background assumptions. The confirmation relation should be construed as a three-place relation between a hypothesis, evidence (from multiple modes), and the various background assumptions required to relate evidence from each mode to the hypothesis. Background assumptions are like any belief: they have varying degrees of plausibility. Some are dodgy. A mode of evidence can provide independent evidence for a background assumption of another mode of evidence. Thus, one evidential

mode can support a background assumption which is necessary to relate evidence from another mode to the hypothesis; of course, the evidential support for the first background assumption will require its own background assumptions. Staley (2004) has argued that this is an important use of multimodal evidence. The background assumptions of a single mode of evidence can themselves be supported by independent evidence. Then, when the first mode of evidence confirms a hypothesis, the support that this evidence provides to the hypothesis is indirectly strengthened by evidence from other modes which support auxiliary assumptions required for the first mode.

This is a compelling and rather straightforward way to construe the value of multimodal evidence. We should be clear about the difference between security and robustness. Security does not require multiple concordant modes of evidence for the *same* hypothesis. After all, security just is the use of one mode of evidence to support an auxiliary hypothesis for another mode of evidence, which is itself evidence for the main hypothesis of interest. Thus, security avoids the challenge of amalgamating multimodal evidence which I discuss in Section 9.6. Indeed, one can gain security simply by using a single mode of evidence for a hypothesis, as long as the auxiliary hypotheses for this mode of evidence are supported by other, independent modes of evidence. One might think that we can construe such an evidential situation as robustness with a single mode of evidence. However, it is helpful to maintain the distinction between robustness and security, since the structure of the arguments are different. Moreover, we should not be misled by diction. Security, presumably, is a matter of degree: if the auxiliary hypotheses of a primary hypothesis are supported by independent evidence, then we might be justified in thinking that our primary hypothesis is 'more secure' than if the auxiliary hypotheses were not supported by independent evidence, but we would not be justified in thinking that our primary hypothesis is 'secure' tout court.

9.4 Rerum Concordia Discors

If concordant multimodal evidence provides greater epistemic support to a hypothesis, it is unclear what support is provided to a hypothesis in the more common situation in which multimodal evidence is discordant. Franklin recently raised the problem of discordance, and suggested that it can be solved by various methodological strategies, which prominently include generating more evidence from independent techniques (2002). While Franklin is correct to identify discordance as a problem for what he calls the "epistemology of evidence", and his appeal to a plurality of reasoning strategies is valuable, I argue below that what he considers a solution to the problem of discordance is better construed as the source of problem.

Discordance is based on both inconsistency and incongruity. Inconsistency is straightforward: Petri dishes suggest x and test tubes suggest $\neg x$. In the absence of a methodological meta-standard, there is no obvious way to reconcile various kinds of inconsistent data. Incongruity is even more troublesome. How is it even possible for evidence from different types of experiments to cohere? Evidence from

different types of experiments is often written in different 'languages'. Petri dishes suggest x, test tubes suggest y, mice suggest z, monkeys suggest $0.8z$, mathematical models suggest $2z$, clinical experience suggests that sometimes y occurs, and human case-control studies suggest y while randomized control trials suggest $\neg y$. To consider multimodal evidence as evidence for the same hypothesis requires more or less inferences between evidential modes. The various 'languages' of different modes of evidence might be translatable into languages of other modes, if one holds the right background assumptions. That is, seemingly incongruous modes of evidence can both be construed as evidence for the same hypothesis given certain background assumptions that relate each mode to the hypothesis. The background assumptions necessary for such translations will have varying degrees of plausibility. If they are not plausible, then it is hard to see how multimodal evidence provides greater epistemic support to a hypothesis than does a single mode of evidence.

For much of the twentieth century, philosophy of science considered idealizations of evidence – Carnap, for example, developed confirmation theory "given a body of evidence e", without worrying about what constitutes a "body of evidence" (1950). In ideal evidential contexts, robustness is a valuable epistemic guide. Real science is almost never in ideal evidential contexts; recent historical and sociological accounts of science have reminded philosophers of the messy details of scientific inquiry. In Section 9.1 I quickly mentioned Galileo, Wegener, Avery, and Ridge as examples of people grappling with discordant multimodal evidence. The following example more richly illustrates the problem, though the example should hardly be needed, since discordance is ubiquitous.

9.4.1 Multimodal Evidence on Influenza Transmission

Epidemiologists do not know how the influenza virus is transmitted from one person to another. The mode of infectious disease transmission has been traditionally categorized as either "airborne" or "contact".[3] A causative organism is classified as airborne if it travels on aerosolized particles through the air, often over long distances, from an infected individual to the recipient. A causative organism is classified as contact if it travels on large particles or droplets over short distances and can survive on surfaces for some time. Clinicians tend to believe that influenza is spread only by contact transmission. Years of experience caring for influenza patients and observing the patterns of influenza outbreaks has convinced them that the influenza virus is not spread through the air. If influenza is an airborne virus, then patterns of influenza transmission during outbreaks should show dispersion over large distances, similar to other viruses known to be spread by airborne transmission. Virtually no influenza outbreaks have had such a dispersed pattern of

[3] I am, of course, greatly simplifying for the sake of exposition.

transmission. Moreover, nurses and physicians almost never contract influenza from patients, unless they have provided close care of a patient with influenza.

Conversely, some scientists, usually occupational health experts and academic virologists, believe that influenza could be an airborne virus. Several animal studies have been performed, with mixed conclusions. One prominent case often referred to is based on an airplane that was grounded for several hours, in which a passenger with influenza spread the virus to numerous other passengers. Based on seating information and laboratory results, investigators were able to map the spread of the virus; this map was interpreted as evidence that the influenza virus was transmitted through the air. More carefully controlled experiments are difficult. No controlled human experiments can be performed for ethical reasons. However, in the 1960s researchers had prisoner 'volunteers' breathe influenza through filters of varying porosity; again, interpretations of results from these experiments were varied, but suggested that influenza could be airborne. Mathematical models of influenza transmission have been constructed, using parameters such as the number of virus particles emitted during a sneeze, the size of sneeze droplets upon emission, the shrinking of droplet size in the air, the distance of transmission of particles of various size, and the number of virus particles likely to reach a 'target' site on recipients. The probability of airborne influenza transmission is considered to be relatively high given reasonable estimates for these parameters.

Even when described at such a coarse grain the various types of evidence regarding the mode of influenza transmission illustrate the problem of discordance. Some scientists argue (using mathematical models and animal experiments) that influenza is transmitted via an airborne route, whereas others argue (based on clinical experience and observational studies) that influenza is transmitted via a contact route. Such discordance demonstrates the poverty of robustness: multiple experimental techniques and reasoning strategies have been used by different scientists, but the results remain inconclusive. A single case does not, of course, demonstrate the ubiquity of discordance; rather, the case is merely meant as an illustration of what is meant by discordance.

If different modes of evidence support contrary conclusions, there is no obvious way to compare or combine such evidence in an orderly or quantifiable way, let alone to compare such a combination of evidence to evidence from a single mode. Philosophers have long wished to quantify the degree of support that evidence provides to a hypothesis. At best, the problem of discordance suggests that robustness is limited to a qualitative notion. And if robustness is a qualitative notion, how should we demarcate robust from non-robust evidence? At worst, the problem of discordance suggests that evidence of different kinds cannot be combined in a coherent way.

One might respond: discordance is not a problem for robustness, since by definition robust evidence is generated when multiple independent modes give the *same* result on the *same* hypothesis. To appeal to discordant evidence as a challenge for robustness simply misses the point. True, but: the problem of discordance is not a knockdown argument against the value of robustness; rather, discordance demonstrates an important constraint on the value of robustness. Robustness, and its

corresponding methodological prescription – get more data! (of different kinds) – is obviously valuable. However, this prescription is not something that scientists need to be told – they already follow this common-sense maxim.

That multimodal evidence is often discordant is an empirical claim. Some might think this a weakness of the above argument. However, the opposite is, of course, also an empirical claim – that multimodal evidence is often concordant – and this is an empirical claim which is false. History of science might occasionally provide examples of apparent concordance, but concordance is easier to see in retrospect, with a selective filter for reconstructions of scientific success. Much history of science tends to focus on the peaks of scientific achievement rather than the winding paths in the valleys of scientific effort – at least, the history of science that *philosophers* tend to notice, like Nye's account of Perrin's arguments for atoms, is history of scientific success. Philosophers have focused on the peaks of scientific success, but the lovely paths of truth in the valleys of scientific struggle are often discordant.

Here is a more prosaic way of stating a related worry. Concordant multimodal (robust) evidence for x is sufficient, but not necessary, for a high probability of x. Now, notice two problems that stem from this vague formulation. First, actually specifying the high probability of x depends on principled methods of quantifying concordance and assessing and amalgamating multimodal evidence, which we lack, and thus, we cannot specify the probability of x. That x even has a high probability is merely an intuition. Second, x might be true despite a failure of robustness, but robustness-style arguments do not tell us what to believe in situations of evidential discordance. Franklin suggests that robustness helps resolve discordant data, but I have argued the converse: discordant evidence diminishes the value of robustness. Epistemic guidance is needed most in difficult cases, when multiple independent techniques produce discordant evidence. In such cases robustness is worse than useless, since the fact of multiple modes of evidence is the source of the problem. Real science is often confronted with the problem of discordance.

9.5 Individuating Multimodal Evidence

One advantage of the term "multimodal" is that we can attempt to determine the basis of evidential diversity by determining what modes of evidence are. In other words, clarity on what a mode of evidence is will give clarity on what multimodal evidence is. Understanding what a mode is can partly be determined by knowing what individuates one mode of evidence from another mode. A mode is a type of evidence, of which there can be multiple tokens. For instance, a case-control study is a particular type of epidemiological study design, which can have multiple (infinite) tokens, or instantiations, of the type: two case-control studies identical in all respects except for the number of subjects in each study would not thereby make for two different types of case-control studies, but rather would make for two different tokens of the same type. At first glance, understanding what modes are seems straightforward. Consider the following:

> We have an intuitive grasp on the idea of diversity among experiments. For instance, measuring the melting point of oxygen on a Monday and on a Tuesday would be the same experiment, but would be different from determining the rate at which oxygen and hydrogen react to form water. (Howson and Urbach 1989, p. 84)

While I share this "intuitive grasp" of what multimodal evidence is, it is surprisingly difficult to specify a more clear definition of multimodal evidence. This difficulty is based on the challenge of determining what the proper form of independence should be between modes of evidence. What *form* of independence between techniques – material? theoretical? probabilistic? – is sufficient to individuate evidential modes? What *degree* of independence between techniques – total? partial? – is sufficient to individuate evidential modes? What criteria should we use to individuate modes of evidence? Individuation of modes of evidence is relative to the intended use of the evidence; several uses of multimodal evidence have been suggested in Sections 9.2 and 9.3. Here I consider the independence between modes necessary for robustness arguments.

The individuation problem can be motivated by considering the following simple case, similar to that in the passage from Howson and Urbach. When testing the efficacy of a drug, we might use chemical assays, animal studies, and human trials, each of which we would intuitively describe as a different mode of evidence, and so this would be a case of multimodal evidence. In contrast, performing a particular animal experiment on one day, and then performing the same experiment with all the same parameters again on another day, would not thereby generate two modes of evidence, and so this would not be a case of multimodal evidence (we could call it a case of *mono*modal evidence). Why does the former set of experiments generate *multi*modal evidence and the latter set of experiments only generate *mono*modal evidence? If we had a criterion for the individuation of modes of evidence then we could answer this question, and we would be far along the way to an adequate understanding of what multimodal evidence is and what conditions must be met in order to make a robustness-style argument.

One suggestion is due to Culp (1994): a necessary condition for robustness-style arguments is that modes of evidence should rely on different background theories. It is a commonplace view that evidence is theory-laden, and Culp's suggestion is that the different modes of evidence in a robustness argument must be laden with different theories. But not all evidence is theory-laden in the same way or to the same degree. And sometimes knowing what theory ladens the data is difficult or impossible. Further, I can imagine two pieces of evidence which depend on the same theory for the production of data and interpretation of evidence, and yet which we would call different modes. Consider, for example, all the possible study designs in epidemiology (case-control studies, cohort studies, randomized controlled trials, and so on). Although each of these modes requires particular background assumptions to relate evidence from the mode to a hypothesis, such background assumptions are not necessarily *theories*, if one pedantically reserves this term for high-level scientific abstractions; perhaps some theory is used in interpreting the evidence from these designs, but they are not necessarily *different* theories which laden the evidence

from different epidemiological study designs; and yet, these study designs are considered to be different modes of evidence by epidemiologists (though of course they do not use my terminology). Moreover, it is easy to imagine a robustness argument based on evidence from multiple epidemiological studies of different designs. The unit of theory is too coarse-grained to serve as a basis of individuation. Individuation of modes needs a finer-grained criterion than theory independence.

Given that all data is only evidence relative to a hypothesis in conjunction with certain background assumptions, another way to conceptualize the individuation of modes of evidence is by the independence of background assumptions between the modes, relative to a given hypothesis. To individuate two modes, it might be sufficient if the modes share all the same background assumptions except one. One might think that this is not restrictive enough. To consider Tuesday's animal experiment as the same mode as Thursday's animal experiment, besides assuming that the animal experiments followed the same protocol, we must hold several background assumptions on Thursday that we didn't on Tuesday – that the bit of nature under investigation has retained its causal structure since Tuesday, that the different socks which the scientist is wearing on Thursday does not influence the results of the experiment, that the change in the moon's gravity does not influence the results of the experiment, and so on – and yet we would not thereby call these animal experiments two different modes of evidence. Thus it is necessary to have at least a few unshared background assumptions between even tokens of the same mode, let alone between multiple modes.

The other extreme of independence of background assumptions would be when two modes are individuated based on a total exclusivity of background assumptions; that is, when the evidential modes do not share a single background assumption. This might also be too restrictive, since one might think that at bottom all modes of evidence, at least when related to the same hypothesis, must share at least *some* background assumptions. Think of the sensory modalities: vision and touch, though seemingly very distinct modes of sensation, rely on much of the same cognitive apparatus.

Since our knowledge of many background assumptions can be far less than certain, our interpretation of almost any data as evidence for a hypothesis might be an artifactual interpretation based on false background assumptions. A robustness argument based on evidence from different modes, with different background assumptions, might be compelling if the *problematic* assumptions for each mode of evidence – those assumptions which we are uncertain about – were different between modes. Consider a situation in which evidence from a case-control study with high external validity and low internal validity is concordant with evidence from a randomized controlled trial (RCT) with high internal validity and low external validity. To think that both modes of evidence are truth-conducive for a general hypothesis of interest (that is, that both modes of evidence give evidence that is true *and* general, or internally *and* externally valid), it is necessary to hold certain background assumptions for each mode. For the case-control study, a required assumption is that there is no selection bias. For the RCT, a required assumption is that the results are exportable to our population of interest. These evidential

modes are individuated rather weakly. They are both human studies at a population level, and as such they share many assumptions, and the statistical analysis of the data from the two modes rely on the same assumptions about population structure. However, the particularly problematic assumptions are the unshared ones. Given that they are unshared, if the two kinds of studies give concordant evidence, that is a reason to think that the unshared background assumptions are not as problematic (in this particular situation) as we would otherwise expect, and so that the evidence is truth-conducive. So *problematic-auxiliary independence* is a good candidate for individuating modes of evidence for arguments based on robustness. The robustness argument for this example would then go as follows. If there was a positive result in the RCT, we might be wrong in assuming that we can generalize its results to a broader population, because of the RCT's low external validity. If there was a positive result in the case-control study, we might be wrong in assuming that the positive result was a true finding, because of the case-control study's low internal validity. But the probability that both studies committed an error is less than the probability that either study committed an error separately.

Thus we can say: it is the background assumptions which we are uncertain about that matter for individuating modes. We can then account for robustness in the following way. A hypothesis is more likely to be true when two or more modes of evidence provide concordant multimodal evidence and the worrisome or problematic auxiliary assumptions for all modes of evidence are independent of each other. At least one problem with attempting to individuate modes based on problematic-auxiliary independence is that we must assume that we can individuate assumptions and determine which assumptions are problematic. This, presumably, can only be done on a case-by-case basis. But how do we know which assumptions are problematic? We could describe the "causal history" or the "mechanism" of a mode of evidence – that is, we could list all the entities and relations involved in the production of the evidence – and then say that if the causal history contains an entity or a relation which is somehow unreliable, then it is the assumptions about that entity or relation which are problematic. This is just pushing the individuation problem back a level: now we have to identify those worrying entities, for which I doubt there is any general criterion of identification. Consider a comparison between electron microscopes and witnesses: evidence from an electron microscope should be construed as being of a different mode than evidence from personal testimony. Two common assumptions thought to be problematic for evidence from personal testimony are based on the witness's capability and the witness's honesty. But a person, the microscope operator, was also involved in the generation of evidence from an electron microscope, and yet we do not normally worry about the capability or the honesty of the microscope operator. It is almost always safe to assume that the microscope operator is honest and capable. Both modes of evidence have, in their causal history, the same type of entity and its associated activity: a person who relays their experience of the world. Despite this similarity, in one mode of evidence the entity has associated problematic assumptions and in the other mode of evidence the entity does not have associated problematic assumptions. Of course, various stories could be told to explain this. My point is that as a criterion of individuation of modes,

appealing to problematic background assumptions shifts the burden from specifying a satisfactory and general criterion of individuating modes to specifying a satisfactory and general criterion of identifying problematic background assumptions. This is a burden unlikely to be met.

The prospect of identifying a general definition of multimodal evidence, based on a criterion of individuation between modes, is more difficult than one might have at first thought. This does not entail that, in fact, there are no modes, or that the difference between multimodal evidence and monomodal evidence is illusory or arbitrary. It just means that drawing a sharp demarcation might be impossible. Nor does this mitigate the epistemic importance of multimodal evidence. After all, there does not exist a compelling criterion to individuate sensory modalities, and yet we assume that there are multiple sensory modalities and that having multiple sensory modalities is epistemically important (Keeley 2002). Same with multimodal evidence: we might not be able to come up with a compelling definition of multimodal evidence based on a criterion of individuation for modes, but multimodal evidence remains profoundly important.

9.6 Amalgamating Multimodal Evidence

I suggested that multimodal evidence is said to be important because it is conducive to both certainty, when the evidence from the available modes is concordant, and to uncertainty, when the evidence from available modes is discordant (Section 9.1). But I also suggested that these views of multimodal evidence – that concordant multimodal evidence is conducive to certainty and that discordant multimodal evidence is conducive to uncertainty – are in themselves unsatisfactory. Metaphors like 'the weight of the evidence' or 'robust results' are usually too vague to warrant assent in the hypothesis in question, and indeed, many scientific controversies are disputes about what the weight of the evidence actually is, or if the results are actually robust or not. If disputants in a scientific controversy had a principled amalgamation function for multimodal evidence, then arguments based on multimodal evidence would be more compelling. Likewise, philosophers making robustness-style arguments would be more convincing if their arguments based on multimodal evidence were supplemented with ways to amalgamate the evidence. Most sciences have crude amalgamation functions for multimodal evidence, but since multimodal evidence is so poorly understood, we have no way to systematically compare or assess the various multimodal evidence amalgamation functions currently in use. I will briefly sketch the contours of what such a function might look like.

To know the impact of multimodal evidence on the confirmation or disconfirmation of a hypothesis, all relevant modes of evidence must be assessed and amalgamated. Modes of evidence should be assessed on several desiderata, including quality, relevance, salience, and concordance. These desiderata have been discussed in detail by others, but to support my argument I will briefly mention them here.[4]

[4] See Galison (1987), Cartwright (2007).

Quality is a straightforward notion which refers to the degree to which a mode is free from systematic errors. *Relevance* refers to the plausibility of the background assumptions that are required to believe that data from a particular mode is evidence for or against a hypothesis. A mode is highly relevant to a hypothesis if data from the mode can be justifiably interpreted as evidence which confirms or disconfirms the hypothesis when such an interpretation requires few implausible auxiliary assumptions. Another important desideratum of evidential assessment is *salience*, which refers to the strength or intensity of results from a mode, or the impact of a unit of evidence on our credence. For example, when testing the efficacy of a new drug to treat depression, if the symptoms in the treatment group improve by five percent compared to the placebo group, that would be a less salient finding than if the symptoms in the treatment group improve by fifty percent compared to the placebo group. Finally, *concordance* is a measure of the degree of consistency of evidence from all the relevant modes for a particular hypothesis. If evidence from all the modes allows for the same inference, given reasonable auxiliary assumptions for each mode, then that multimodal evidence is concordant. Quality, relevance, salience, and concordance do not exhaust the important evidential desiderata, but they are among the most important features of evidence.

Scientists lack systematic methods for assessing quality, relevance, salience, and concordance, though some disciplines have criteria for determining what counts as high quality evidence. For example, the evidence-based medicine movement rank-orders various kinds of studies, with evidence produced by RCTs considered the highest quality of evidence; evidence from prospective cohort studies, case-control studies, observational studies, and case reports normally follow RCTs in descending order of quality.

Different modes of evidence and combinations of modes will satisfy the desiderata to various degrees in different circumstances, by various amalgamation functions. Part of what a good multimodal evidence amalgamation function should do is assess multiple modes of evidence on these multiple evidential criteria: each mode of evidence must be assessed on its quality, relevance, and salience, and the set of the modes of evidence together should be assessed on its concordance. The basis of many scientific controversies can be construed as disputes about differential assessments of these desiderata: one group of scientists might believe that evidence from some techniques is of higher quality or is more relevant to the hypothesis or has greater confirmational salience than other techniques, while another group of scientists might believe that evidence from the latter techniques is of higher quality or is more relevant or salient. For example, Galison argues that one tradition in particle physics considers an image of a "golden event" to be compelling evidence – an observation of a golden event provides strong confirmation to a hypothesis; whereas another tradition in particle physics considers repeatable observations on which statistical analyses can be performed to be compelling evidence.

Abstractly, an amalgamation function for multimodal evidence should do the following: evidence from multiple modes would be fed into the amalgamation function, which would assess evidence on prior criteria (quality of mode), relative criteria (relevance of mode to a given hypothesis) and posterior criteria (salience of

evidence from particular modes and concordance/discordance of evidence between modes), and the output would be a constraint on our justified credence. The construction and evaluation of such schemes should be a major task for theoretical scientists and philosophers of science. There currently are functions that combine quantitative evidence from different modes and have a quantitative output, including Demspter-Shafer Theory, Jeffrey conditionalization, and statistical meta-analyses, and there are functions that combine qualitative evidence from different modes and have a qualitative output, including narrative synthesis, meta-ethnography, and there are functions that combine quantitative evidence from different modes but have a qualitative output, such as the evidence hierarchy schemes in evidence-based medicine. An investigation into the methodological virtues and constraints of these functions would be interesting (for example, Stegenga (2011) assesses the purported merits of meta-analysis). With such amalgamation functions, robustness-style arguments might then be more compelling, because there would be a systematic way to guide credence when presented with multimodal evidence. Such functions would be especially valuable when multimodal evidence is discordant. The extent to which robustness-style arguments could be made might be increased if they could be based on multimodal evidence which is not concordant.

9.7 Conclusion

One of the ways that multimodal evidence is said to be valuable is robustness: that is, when multimodal evidence for a hypothesis is concordant, that hypothesis is more likely to be true, or explanatory, or phenomena-saving, or whatever predicate of epistemic success fits most comfortably with one's philosophical inclinations. I have raised several challenges for robustness, the most prominent of which is the ubiquity of discordance. Despite idealizations of scientific success, the world is usually a *rerum concordia discors*. Without the use of compelling schemes to amalgamate discordant multimodal evidence, robustness arguments are vague. Amalgamation functions could provide more constraint on our justified belief in a hypothesis when presented with multimodal evidence.

Appendix: Bayesian Amalgamation

Here I briefly consider how one might amalgamate evidence using a Bayesian approach. Bayesian conditionalization is a rule for revising one's probability of a hypothesis upon receiving evidence. If a scientist learns e, and $p_{\text{old}}(H)$ is the scientist's assessment of the probability of a hypothesis *before* receiving the evidence, then $p_{\text{new}}(H)$ – the scientist's assessment of the probability of the hypothesis *after* receiving the evidence – should equal $p_{\text{old}}(H|e)$. Since this latter term is a conditional probability, it can be calculated using Bayes' Theorem (BT):

$$\text{(BT)} \quad p(H|e) = p(e|H)p(H)/p(e)$$

This suggests a possible way to amalgamate multimodal evidence, based on what is sometimes called 'strict conditionalization' (SC): we could update the probability of the hypothesis by sequentially conditionalizing with Bayes' Theorem for each mode of evidence.[5]

$$\text{(SC)} \quad p_{\text{new}}(H) = p_{\text{old}}(H|e) = p(e|H)p_{\text{old}}(H)/p(e)$$

One could arbitrarily order available modes from 1 to n, and then use Bayes' Theorem to update the probability of the hypothesis sequentially for each mode, and the posterior probability of the hypothesis after updating on evidence from mode n would become the prior probability of the hypothesis for updating on evidence from mode n+1. The probability of the hypothesis after conditionalizing on the evidence from the first mode would be as above, substituting numerical subscripts for evidence from each mode in place of 'old' and 'new':

$$p(H|e_1) = p(e_1/H)p(H)/p(e_1)$$

The posterior probability, $p(H|e_1)$, would then be the 'new' prior probability, $p(H)$ for updating by evidence from the next mode, e_2:

$$p(H\,|e_2\,) = p(e_2|H\,)p(H|e_1\,)/p(e_2)$$

This sequential updating would continue until the evidence from the final mode, n was used to update the penultimate probability of the hypothesis $p\big(H_{f-1}\big)$ to determine the final probability of the hypothesis $p(H_f)$:

$$p\big(H_f|e_n\big) = p(e_n|H\,)p\big(H_{f-1}\big)/p(e_n)$$

Some Bayesians might consider this approach to be the best way to amalgamate multimodal evidence. Several conditions must be met for this method of sequential conditionalization. For all modes of evidence, all terms in Bayes' Theorem must be known: that is, for all modes i, $p\,(e_i|H\,)$ must be known; the initial $p(H)$ must be known (this condition has generated much worry, known as the 'problem of the priors'); and for all modes i, $p(e_i)$ must be known. Determining these terms in practice is often impossible. Consider again the evidence presented in Section 9.4 regarding influenza transmission. What was the probability of observing the pattern of influenza transmission on the landed airplane, conditional on the central competing hypotheses, for example? What was the prior probability of the Contact hypothesis? What was the prior probability of the Airborne hypothesis? Now repeat these questions for the other hypotheses and modes of evidence.

[5] Dutch Book arguments are meant to show that one is rationally required to use SC to learn from evidence.

Also troubling is that Bayes' Theorem requires the scientist using the theorem to know e to be true once e has been observed. In most scientific contexts this is unrealistic. Consider an example given by Skyrms (1986): suppose I see a bird at dusk, and I identify it as a black raven, but because of the evening light, I do not hold the proposition "the bird is a black raven" as my evidence e with perfect confidence (that is, $p(e) \neq 1$). Rather, I might believe e to be true with probability 0.95. Jeffrey (1965) proposed a modification of Bayesian conditionalization to deal with cases in which evidence is uncertain (which, it is reasonable to suppose, is wholly ubiquitous in science). Jeffrey conditionalization (JC), sometimes referred to as 'probability kinematics', is as follows: given multimodal evidence e_i one's updated probability in H, $p_{\text{new}}(H)$, should be:

$$\text{(JC)} \qquad \forall i_{1-n} \sum p_{\text{old}}(H|e_i)\, p_{\text{new}}(e_i)$$

In other words, this is a weighted average of strict conditionalization. To use JC for amalgamating multimodal evidence, one would sequentially update the probability of the hypothesis using JC, similar to the sequential procedure used with SC.

Bayesianism is beset with many well-known problems. This is not the place to rehearse them. But are there any problems with Bayesianism that arise specifically with respect to amalgamating multimodal evidence? A condition of the particular method described above was an arbitrary ordering of the modes. Whatever ordering is chosen should not affect the final probability of the hypothesis. Unfortunately, it is a common complaint against JC that it is 'non-commutative' – the order in which particular pieces of evidence are used to update the probability of the hypothesis makes a difference to the final probability of the hypothesis (see van Fraassen 1989). This problem could be mitigated if there were a way of ordering modes which was superior to others. One might think that if we ordered modes by quality, and used JC on the highest quality mode first and subsequently conditionalized on modes in decreasing order of quality, then the non-commutative property of JC would at least be minimized, because evidence from lower quality modes ought to have a lower impact on the hypothesis anyway. The trouble with this approach is that, despite what some have claimed in particular domains such as evidence-based medicine, there is no general, decontextualized way to order modes of evidence according to a unitary desideratum such as quality. Any ordering of modes will be arbitrary in some important respect. Thus, one cannot resolve the non-commutativity of JC in this way.

Acknowledgments I am grateful for detailed feedback from Léna Soler, Emiliano Trizio and members of the UCSD Philosophy of Science Reading Group, especially Nancy Cartwright. I also benefited from discussion with audiences at the 2008 Canadian Society for the History and Philosophy of Science conference, the 2008 Philosophy of Science Association conference, and the 2008 workshop on robustness hosted by the Archives Henri Poincaré, Laboratoire d'Histoire des Sciences et de Philosophie (Nancy-Université).

References

Achinstein, Peter. 2001. *The Book of Evidence*. New York: Oxford University Press.

Allamel-Raffin, Catherine. 2005. "De l'intersubjectivité à l'interinstrumentalité: L'exemple de la physique des surfaces." *Philosophia Scientiae* 9(1):3–31.

Bechtel, William. 2006. *Discovering Cell Mechanisms*. Cambridge: Cambridge University Press.

Box, George. 1953. "Non-Normality and Tests on Variances." *Biometrika* XL:318–35.

Carnap, R. 1950. *Logical Foundations of Probability*. Chicago, IL: University of Chicago Press.

Cartwright, Nancy. 1983. *How the Laws of Physics Lie*. Oxford: Clarendon Press.

Cartwright, Nancy. 2007. *Hunting Causes and Using Them*. Cambridge: Cambridge University Press.

Chang, Hasok. 2004. *Inventing Temperature*. New York: Oxford University Press.

Collins, Harry. 1985. *Changing Order: Replication and Induction in Scientific Practice*. Chicago: University of Chicago Press.

Culp, Sylvia. 1994. "Defending Robustness: The Bacterial Mesosome as a Test Case." In *Philosophy of Science Association 1994*, vol. 1, edited by David Hull, Micky Forbes, and Richard M. Burian, 46–57 Chicago.

Douglas, Heather. 2004. "The Irreducible Complexity of Objectivity." *Synthese* 138(3):453–73.

Fitelson, Branden. 1996. "Wayne, Horwich, and Evidential Diversity." *Philosophy of Science* 63:652–60.

Franklin, Allan. 2002. *Selectivity and Discord: Two Problems of Experiment*. Pittsburgh, PA: University of Pittsburgh Press.

Franklin, Allan, and Colin Howson. 1984. "Why Do Scientists Prefer to Vary Their Experiments?" *Studies in the History and Philosophy of Science* 15:51–62.

Galison, Peter. 1987. *How Experiments End*. Chicago: University of Chicago Press.

Hacking, Ian. 1983. *Representing and Intervening*. Cambridge: Cambridge University Press.

Hempel, Carl. 1966. *Philosophy of Natural Science*. Englewood Cliffs, NJ: Prentice-Hall.

Horwich, Paul. 1982. *Probability and Evidence*. Cambridge: Cambridge University Press.

Howson, Colin, and Peter Urbach. 1989. *Scientific Reasoning: The Bayesian Approach*. La Salle, IL: Open Court.

Hudson, Robert G. 1999. "Mesosomes: A Study in the Nature of Experimental Reasoning." *Philosophy of Science* 66(2):289–309.

Jeffrey, Richard. 1965. *The Logic of Decision*. Chicago: University Of Chicago Press.

Keeley, Brian. 2002. "Making Sense of the Senses: Individuating Modalities in Humans and Other Animals." *The Journal of Philosophy* 99(1):5–28.

Kosso, Peter. 2006. "Detecting Extrasolar Planets." *Studies in History and Philosophy of Science* 37:224–36.

Levins, Richard. 1966. "The Strategy of Model Building in Population Biology." *American Scientist* 54:421–31.

Mayo, Deborah. 1996. *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

Oreskes, Naomi. 1999. *The Rejection of Continental Drift*. New York: Oxford University Press.

Rasmussen, Nicolas. 1993. "Facts, Artifacts, and Mesosomes: Practicing Epistemology with the Electron Microscope." *Studies in History and Philosophy of Science* 24(2):221–65.

Salmon, Wesley. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.

Skyrms, B. 1986. *The Dynamics of Rational Deliberation*. Cambridge, MA: Harvard University Press.

Staley, Kent. 2004. "Robust Evidence and Secure Evidence Claims." *Philosophy of Science* 71:467–88.

Stegenga, Jacob. 2011. "Is Meta-Analysis the Platinum Standard of Evidence?" *Studies in History and Philosophy of Biological and Biomedical Sciences* 42:497–507.

Stegenga, Jacob. 2011. “The Chemical Characterization of the Gene: Vicissitudes of Evidential Assessment.” *History and Philosophy of the Life Sciences* 33:103–26.

Trout, J.D. 1993. “Robustness and Integrative Survival in Significance Testing: The World’s Contribution to Rationality.” *British Journal for the Philosophy of Science* 44:1–15.

van Fraassen, Bas. 1989. *Laws and Symmetry*. New York: Oxford University Press.

Weber, Marcel. 2005. *Philosophy of Experimental Biology*. Cambridge: Cambridge University Press.

Westman, Robert. 2011. *The Copernican Question: Prognostication, Skepticism, and Celestial Order.* Berkeley, CA: University of California Press.

Wimsatt, William. 1981. “Robustness, Reliability, and Overdetermination.” In *Scientific Inquiry and the Social Sciences*, edited by M. Brewer and B. Collins, 124–63. San Francisco, CA: Jossey-Bass.