Penultimate draft (July 2012)

Forthcoming in *Philosophy of Science.*

Forging Model/World relations: Relevance and Reliability

Isabelle Peschard

Department of Philosophy, San Francisco State University.

Abstract: The relation between models and the world is mediated by experimental procedures generating data that are used as evidence to evaluate the model. Data can serve as empirical evidence, for or against, only if they result from *reliable* experimental procedures. The aim of this paper is to discuss the role of relevance judgments in the evaluation of reliability and to clarify the conditions under which reliability can be a strictly empirical matter. It is argued that reliability is a strictly empirical issue only in the restricted case where the claim under test/investigation is about a data-generating procedure.

The author wishes to acknowledge support for this research by NSF grant SES-1026183

Forging Model/World relations: Relevance and Reliability

1. Introduction. The relation between models and the world is mediated by experimental procedures. Experimental procedures achieve this mediation through the generation of data used as evidence to evaluate the models. Data however can serve as empirical evidence, for or against, only if they result from *reliable* experimental procedures. So how to understand the relation between models and the world will depend on what it takes for an experimental procedure to be reliable. A view of experimental reliability as strictly empirical matter (Bogen & Woodward 2003), for instance, might be used to support a conception of empirical justification as independent of value judgments. It is the aim of this paper to discuss whether and how experimental reliability can be strictly an empirical matter.

Consider this excerpt from an article discussing the cost of crime in Nova Scotia[1]:
"… a substantial increase in common assaults has raised the official violent crime rate in the province above the national average. *A substantial portion of the increase in the official crime rate is due to higher reporting rates for assaults, sexual assaults, domestic violence* and other crimes, a positive sign signifying reduced social tolerance for violent behavior once considered socially 'acceptable'." [italics added]

What the official violent crime rate is might be an empirical judgment, strictly based on measurement, but it clearly depends on the prior, not strictly empirical, judgment, call it a 'judgment of relevance', about what qualifies, and should be counted, as violent crime.[2] As we will see, there are good reasons to think that such judgments might play a role in the evaluation of the reliability of measurement procedures not just in social sciences but in natural sciences as well. Some conceptual clarification is needed to see how that can be so. That will be done in sections 2 and 3. Section 4 contrasts this view of a role played by such relevance judgments with a conception of reliability as strictly empirical. Section 5 discusses an episode in particle physics that illustrates strictly empirical reliability but also undermines the pertinence of the contrast of social vs. natural phenomena. It suggests that the existence of a theoretical framework might be a more pertinent factor. Section 6 however shows that this account also can easily be undermined and illustrates the point with a case study in neuroscience. Section 7 finally proposes an account of the conditions in which experimental reliability can be strictly empirical. The originality of this account is that it is in terms of the type of claim under test/investigation: for reliability to be strictly empirical the claim under test/investigation must be about a data-generating procedure. But that is far from always being the case.

---

1 Ronald Colman, 'The cost of crime in Nova Scotia.' *Journal of Business Administration and Policy Analysis* (January 1999) In Mark Battersby (2010: 62)

2 That domestic violence should count as crime is of course not sufficient to account for the increase in *reporting* rate.

2. Conceptual clarification. Relevance judgments, in the context of this paper, are judgments about what quantity should be measured or, correlatively, what effect should be taken into account by a measurement procedure. For our purpose a clarification by example and exclusion together with a very general characterization will suffice. By example: the judgment that domestic violence should be counted as violent crime is a relevance judgment. It takes the rate of domestic violence to be relevant to the evaluation of the official crime rate. By exclusion: relevance judgments are not empirical judgments, even if they may appeal to some empirical considerations[3]. They can be, in general, characterized as normative judgments.

Reliability can be attributed to claims, procedures, or data. These attributions are closely related: unreliable measurement procedures produce unreliable data and unreliable data lead to unreliable claims. The discussion that follows will be limited to the reliability/unreliability of measurement procedures.

To say that the procedure is reliable is tantamount to saying that it is capable of producing evidence for or against a given claim. As will be seen, a measurement procedure is not reliable or not reliable *per se* but with respect to a given claim. More on this point in the next sections. For now let's explicate further some aspect of the notion of reliability. Suppose that H is the hypothesis under evaluation and M a measurement procedure producing data that agree with H. The procedure M is reliable for evaluating H only if it is very unlikely that it would produce data that agree (or agree so well) with H if H was false. This reliability requirement is adapted from Deborah Mayo's notion of severity. According to Mayo (1996, 2000), a hypothesis H passes a severe test T with data D if and only if the data D fit H (for a suitable notion of fit or 'distance') and the test has a very low probability of producing a result that fits H as well as (or better than) D, if H were false or incorrect. I read the fulfillment of this later condition as a reliability requirement on the procedure that produces the data: the procedure must be reliable in that it should be very unlikely to lead us astray in producing data that look like supporting evidence for a claim that is false.

It will prove more convenient for the discussion to use the criterion for unreliability that the reliability requirement entails:

Unreliability criterion:

> A measurement procedure M producing data that agree with H is not reliable if it can be shown that this procedure might well have produced results that agree as well with H even if H were false.

To see how the criterion works, let's apply it to the crime example introduced earlier. The application requires specifying the procedure M, producing the data, and the hypothesis H that is under test. The

[3] Relevance judgments, as they are understood, are not to be confused with judgments which, in a mechanistic account of explanation, state that a certain factor is causally or constitutively relevant. In this latter case, such a judgment states the causal contribution of this factor to the evolution of the variable of interest. In this paper, as will become clearer, what is at stake in a relevance judgment is whether the causal contribution of certain factors should be taken into consideration at all.

reformulation of the example will also provide us with a formal template appropriate to a general discussion of reliability.

3. The crime rate example reformulated. Simplifying the case a little for the sake of the argument the reformulation of the crime rate example given in the previous section goes as follows:

- The phenomenon P: crime rate evolution
- Claim/hypothesis about the phenomenon that is under test:
  Ho(P) : "There is no increase in crime rate"
- Two measurement procedures:

M1 (O): measurement only of the rate of outdoor aggressions and produces data O.
M2 (O, D): measurement of the rate of outdoor aggressions (data O) and the rate of domestic violence (data D).

Let's assume that the outcome of the procedure M1(0) shows no increase in the rate recorded, in agreement with Ho(P), and hence, provides apparent evidence *for* Ho(P); and that the outcome of the procedure M2(0, D) provides apparent evidence for an increase in crime rate, *against* Ho(P). The data produced by a procedure are evidence for or against only if the procedure is reliable. Let's consider the procedure M1(O) and ask whether it is a reliable procedure.

The answer depends on whether data D for domestic violence should be used to assess the claim about P. If data D should be included, then the results produced by M1(O) could agree with Ho(P) as well as they do even if Ho(P) is false and there is, in fact, an increase of the crime rate. It does not matter whether the data D actually does or does not affect the crime rate. What matters is that *it could* and that M1(O) would be insensitive to this influence. So we would have to conclude that M1(O) is not a reliable procedure. If data D should *not* be included, then M1(O) might be a reliable procedure. We would need to know more about M1(O) to reach this conclusion.

It is certainly, at least in part, on the basis of empirical considerations that one would argue for including D and treating domestic violence as violent crime. One might, for instance, provide empirical evidence of psychological or physical consequences of domestic violence. But whether, given these or other considerations, it should or should not be included as violent crime is not itself an empirical judgment. There must be a decision as to whether these considerations provide sufficient reason to count domestic violence as crime, and the rate of domestic violence as relevant to the evaluation of crime rate. For this reason, the reliability of the measurement procedure M1(O) is not a strictly empirical matter but depends on what we will call 'relevance judgments'. The conclusion about this case then is that such non-empirical judgments play a critical role in the evaluation of reliability.

4. Reductionist view on reliability. The conclusion of the previous section stands in stark contrast to the idea, defended by James Bogen and James Woodward, that "whether or not a detection process is generally reliable is always an empirical matter" (2003:237) and that it is "highly specific empirical facts about the general reliability of particular methods of data production and interpretation … that

are relevant to determining whether data are evidence for certain claims about the phenomenon" (2003:239).

Several explanations might be proposed to explain this contrast. The simplest explanation would be a difference in what the term 'reliability' is targeting. But it is also the easiest to put aside because, for Bogen & Woodward, a generally reliable process is one that has a high probability to discriminate correctly among a set of relevant alternatives (2003:237) and a type of procedure that is not unlikely to produce apparent support for a claim that is false would not satisfy this condition.

Another suggestion might be that Bogen & Woodward's emphasis on the empirical dimension of reliability was as much an emphasis on the inadequacy of a logical account of evidential reasoning. "It is", they write, "specific empirical facts… *and not the formal relationship* that are relevant to determining" evidential support. Nothing however is further from the aim of this paper than a strictly logical account of evidence and reliability. The issue is rather whether, assuming that indeed evidential reasoning is not a merely formal matter, it has to be a strictly empirical matter.

One may also be tempted to think that the lesson from the crime rate example is of limited scope. It concerns, at best, social phenomena but does not support any conclusion regarding natural phenomena. One may even argue that the role played, in this case, by relevance judgments in the evaluation of the procedural reliability is precisely a symptom that the claim under test is about a social phenomenon. Relevance judgments may have some role to play in social sciences, but in natural sciences, one may go on, reliability is just an empirical affair. The next case study, in particle physics, will help to evaluate this objection.

5. Parity conservation. By contrast to the crime rate example examined in the previous section, the case study examined in this section pertains to natural science. The aim is to cast some light on what it takes for reliability to be an empirical matter and on the pertinence, with respect to this issue, of the contrast between natural and social phenomena.

Alan Franklin (1989) views the scientific experiments that provided evidence for the violation of parity conservation as one of the clearest cases of crucial experiments for they overthrew what was "a widely and strongly held belief of the physics community from about 1927 to 1957" (p. 22). At the basis of this belief in parity conservation was an inclination towards the idea of a "general principle of right-left symmetry of Nature". But the claim that parity is conserved was taken to be empirically supported by experimental results from strong and electromagnetic interactions and, interestingly, these results will not be called into question. As they put forth the hypothesis that parity conservation might be violated, Lee and Yang recognized these results. But they also found, "to their surprise", that no experimental support for conservation of parity came from weak-interaction physics. It is on the basis of later results from experiments in weak interaction physics that the belief in parity conservation will be overthrown.

As in the previous section with the crime rate example, there will thus be apparent conflicting evidence, this time depending on whether results from weak interaction are taken into account or not. Let's see what happens when we ask whether the procedure producing the results apparently in favor

of parity conservation is reliable. We first apply the same formal template as the one used for the crime rate example:

- Phenomenon P: parity conservation
- Claim about the phenomenon that is under test:
  Ho(P) : "There is conservation of parity"
- Two measurement procedures:

M1(S): measurement made only for strong interactions (S) – produces apparent evidence *for* Ho(P)
M2(S, W): measurement made for strong interactions (S) and for weak interactions (W) – produces apparent evidence *against* Ho(P)

For the sake of simplicity, only two sorts of interactions are considered, strong and weak interactions; M1(S) represents as one procedure the set of all the experiments with strong interactions that produced apparent evidence for Ho(P); and M2(S, W) represents the experiments with strong and weak interactions.

The procedure M1(S) is reliable only if it is very unlikely that it would produce results that agree with Ho(P) if Ho(P) was false. So if the results from measurements in weak interactions fall under the scope of the claim about parity then M1(S), which is not sensitive to what happens in this type of interactions, is not reliable. For it is not unlikely at all that it would produce results that fit Ho(P) even if Ho(P), because of the violation of parity in weak interactions, is false.

What is at issue is whether the reliability/unreliability of M1(S) is a strictly empirical affair, that is, based strictly on the results of an empirical investigation. Such an empirical investigation would consist in various manipulations and comparisons to test the good functioning of the instruments or check for the possibility of interfering factors, or even to evaluate the statistical significance of the results. The aim of these various techniques is to establish that the procedure in question is capable of measuring, with an appropriate degree of accuracy, what it is intended, by the experimenter, to be measuring. The procedure needs to be (causally) sensitive to whatever factor it is intended to be sensitive to and not, or only minimally, sensitive to some other factors (unless the effect of these other factors can be estimated and subtracted). Let's admit that the judgment to the effect that this condition is satisfied, based on the application of various techniques of, calibration, replication of the results, causal analysis, statistical analysis, selective interventions and manipulations, etc., is an empirical judgment[4]. If that was *all* there is to reliability, then it seems that M1(S) should be deemed reliable. For there was no doubt that M1(S) was indeed measuring exactly what it was intended to measure: some clearly specified effects happening in strong interactions; and was doing it, as Lee and Yang wrote at the time, "to a high degree of accuracy".

The reason why the data produced by M1(S) were not actually regarded as evidence for parity conservation is rather that whatever M1(S) was supposed to detect was not regarded as sufficient ground for the claim in question. Part of the reason for this insufficiency is that M1(S) is insensitive

[4] This assumption might be objected to on the basis, for instance, that judgments of relevance enter in the determination of which possibly interfering factors should actually be tested (Wheeler 2000) or in the determination of some criteria for the test (Staley 2005).

to what happens in weak interactions and that it is thus insensitive is an empirical judgment. But another part must be the judgment that only a procedure sensitive *both* to strong and weak interactions can produce a sufficient ground for parity conservation. And this judgment looks like a relevance judgment.

If the reliability of M1(S) depends on the relevance of results both from strong and weak interactions, we seem to reach, for the same reason, the same conclusion as in the crime rate example: reliability is not strictly an empirical matter.

There is, however, some reason to refuse the analogy with the crime rate example: whereas there was room for debate about whether 'crime' includes domestic violence, there was no doubt that a claim about parity conservation had to be based on results from both strong and weak interactions. There was no need to make a judgment of relevance. That results from weak interactions had to be taken into account seems simply to follow from what 'there is conservation of parity' meant all along: conservation of parity in ALL forms of interaction.

In consequence, the judgment that the procedure limited to measurement in strong interactions was not reliable now appears to be an plain empirical judgment: the judgment to the effect that the procedure was simply not doing what it was supposed to do given a proper theoretical understanding of the claim under test.

What makes reliability a strictly empirical matter in the parity conservation case, by contrast to the crime rate case, is not, as we can see, a difference between natural vs. social phenomena. It is that, in the parity conservation case, the claim under test specifies the procedure required to test it: this procedure should include measurements in all forms of interaction. This specification is provided by a theoretical framework that specifies the procedural interpretation of the concepts used to formulate the claim, their interpretation in terms of measurement procedures. That suggests an alternative explanation of why there might be sometimes room for relevance judgments in the evaluation of reliability: the absence of a theoretical framework that specifies this procedural interpretation and, thereby, what data are relevant.

The problem, though, with this account in terms of presence or absence of a theoretical framework, is that, as will be shown in the next section, the existence of a theoretical framework is actually far from being sufficient to guarantee that reliability will be a strictly empirical matter.

6. Receptive Field. The previous sections led us to associate the role played by relevance judgments in evaluating the reliability of a test to the absence of a theoretical framework that specifies the testing procedure. That includes cases where there is not one theoretical framework that uniquely specifies what data are relevant. A striking illustration of this type of situation is found in the study of cognition and the frustration of some researchers with regards to symbolico-computionalist models of cognition. For those promoting an embodied enactive approach to cognition, for instance, the acquisition of practical, bodily abilities via perceptually guided interaction with the environment is essential to cognition. They see it as a source of relevant data which they think cannot be accounted for strictly in computationalist terms (Varela 1991; Noë 2005).

Another example is offered by Helen Longino with the study of aggressive and sexual behaviors using different theoretical frameworks (Longino 2006). The reliability of the procedure producing data that serve as evidence for any account of gender role behavior will depend on what count as male or female- typical behaviors that have to be accounted for, that is, what features are deemed relevant for the classification (e.g., level of expenditure of physical energy); for some of them, like the hormonal account, it will also depend on the relevance of non-human behavioral data and on whether, since the brain is supposed to mediate the causal action of gonadal hormones on behaviors, the effect of social factors on brain development is deemed relevant and needs to be taken into account. Longino (1990) sees the judgment that non-human behavioral data are relevant as the expression of prior commitment to biological determinism, which itself might result from an even prior commitment to gender dimorphism that makes it possible for biological determinism to be seen as the right conception of biological functioning (Rooney 1992).

A theoretical framework is generally not sufficient to specify completely what data are relevant to the empirical study of a phenomenon. That has much to do with the fact that theories provide principles that guide, and to some extent constrain, but do not determine the construction of their models (Morrison 2007). For some of the factors that have a causal effect on some variable of interest, it may be a source of long-standing controversy whether this effect should be regarded as interference and neutralized or should be regarded as relevant and systematically explored and recorded (Peschard 2011). And, of course, this question might also arise when there is no overarching theoretical principle. A telling illustration of what might be involved in determining what count as relevant data is offered by the neurophysiological investigation of visual cells' receptive field.

Experimental procedures studying the response of the visual cells used to be conducted in an artificial environment with isolated luminous stimuli and an anesthetized animal. The behavior of neurons was then regarded as a characteristic response to a specific type of luminous stimuli. Research over the past decades has shown that this response is sensitive to factors that are absent from the artificial environment. In the light of these experimental findings, the conception of what had to be modeled as receptive field, what data need to be accounted for, has changed. For instance, some studies investigate the effect of the complexity of the environment on the visual cells' response, using, in fact, images as stimuli and statistical analysis to measure their complexity; others emphasize the explorative dimension of visual activity, using a mobile robot equipped with a camera to acquire evidence for it in a 'real' environment; others explore the interpretive dimension of perception, for instance in terms of neural correlates.

All agree that if the earlier procedures are still reliable to test claims about what happens in artificial, passive and impoverished conditions, they only provide a reliable empirical basis to make claims about what is now called the 'classical' receptive field. The reason is that they are not sensitive to factors that are now taken to be relevant to the evaluation of such claims.

The crucial point for us, however, is that there is no general agreement on what these factors are (Chirimuuta & Gold 2009). Different perspectives specify differently the procedural

interpretations and the norms for the empirical evaluation of experimental reliability. But adopting one option over another, and accordingly deciding what data need to be accounted for, is not making an empirical judgment, even if it may be based, in part, on empirical considerations.

7. Data-generating procedure vs. phenomenon. The previous sections have shown that the evaluation of experimental reliability might involve (non-empirical) relevance judgments. No satisfactory account has yet been found of what prevents that from happening or, on the contrary, makes it possible. In this section, an account will be proposed. Instead of looking at the phenomenon under study or the theoretical background, this account focuses on the claim itself that is under test: A measurement procedure is not reliable or unreliable per se but with respect to a given claim and what it takes for an experimental procedure to be reliable depends of the type of claim that is under test.

In order to characterize this dependence relation, it is helpful to think of the experimental procedure as a data-generating procedure. This procedure may be used to evaluate claims about the data-generating procedure itself or claims about some phenomenon under investigation. Whether reliability is or is not a strictly empirical matter depends on whether the claim under test/investigation is or is not of the former type. The link between the claim and the data is the model of the experiment (Suppes 1962). A model of the experiment specifies a set of all the possible outcomes of a specific type of data-generating procedure. More specifically, given a claim under test, the model of the experiment specifies a type of procedure able to produce data that qualify as benchmark for the evaluation of this claim.

Suppose the claim under test is a claim about the relation between two measurable quantities X and Y. This claim can be construed as a claim about a data-generating procedure: one that generates the values of X and Y. As such, this claim itself prescribes a model of the experiment and the corresponding data generating procedure: it should measure the value of Y for different values of X. If the procedure conforms to the prescription of the model of the experiment, it will measure what it is supposed to be measuring and will satisfy the reliability requirement. Whether it does conform to this prescription is an empirical matter investigated via different techniques (calibration, causal analysis and manipulation). What makes, in cases like the one about conservation of parity, the reliability of the procedures an empirical matter is the interpretation of the claim tested as being actually about a data-generating procedure: a procedure that needed to be carried out for all types of interactions. The procedures that neglected to probe weak interactions simply did not conform to the model of the experiment prescribed by the theoretical interpretation of the claim. More generally, then, when the claim under test can be read as a claim about a data-generating procedure, the reliability of the procedure conducted for testing the claim is an empirical matter.

But suppose now that the claim under test cannot be interpreted as being about a data-generating procedure. The issue then arises of what data-generating procedure is adequate for the test and generates data that qualify as benchmark for the evaluation of the claim. It arises with claims about crime rate or about the receptive field. The reliability of a given experimental procedure will depend on the answer to this question. This answer is a judgment that specifies what sort of data need

to be acquired and will need to be accounted for by a model about the crime rate or the receptive field. It is not an empirical judgment; it is what we have called a 'relevance judgment'.

By specifying the appropriate kind of data, relevance judgments transform the claim about, say, the crime rate phenomenon into a claim about a data-generating procedure. And the strictly empirical issue of reliability attached to that latter type of claim will now arise. The distinction between these two sources of unreliability is easily overlooked in discussion of reliability when the claim under test is ambiguously presented as a claim about a phenomenon. This possible ambiguity suggests a new explanation of Bogen & Woodward's claim that reliability is a strictly empirical matter: their claim is restricted to claims about phenomena that can be read as being about a data-generating procedure.

There is a good reason for being, as philosophers, more interested in discussing claims about phenomena than claims about data generating procedures: we are interested in science as an activity of investigation of phenomena, not of data-generating procedures. But whereas any claim about a data-generating procedure can be seen as a claim about some phenomenon, the reverse is not true: not any claim about a phenomenon will be ipso facto a claim about a specific data-generating procedure. Claims about phenomena that cannot (not yet) be read as claims about data-generating procedures are not marginal in science: they are often at the starting point of the modeling of phenomena and at the core of long standing controversies. In such cases, non-empirical judgments of relevance will play a role in the evaluation of the reliability of the testing procedure and, thereby, in forging the relation between models and phenomena.

REFERENCES

Battersby, Mark. 2010. *Is That a Fact? A Field Guide to Statistical and Scientific Information*. Toronto: Broadview Press.

Bogen, James and James Woodward. 2003. "Evading the IRS." *Poznan Studies in the Philosophy of the Sciences and the Humanities* XX: 223-256

Chirimuuta, Mazviita and Ian Gold. 2009. "The Embedded Neuron, The Enactive Field?" In *Oxford Handbook of Philosophy and Neuroscience*, ed. John Bickle, 200-25. Oxford: Oxford University Press.

Franklin, Allan. 1989. *The Neglect of Experiments*. Cambridge: Cambridge University Press.

Longino, Helen. 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton: Princeton University Press

--- 2006. "Theoretical Pluralism and the Scientific Study of Behavior." In *Scientific Pluralism*, ed. Stephen H. Kellert, Helen H. Longino and C. Kenneth Waters. 102-131, Minneapolis: University of Minnesota Press.

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

--- 2000. "Experimental Practice and An Error Statistical Account of Evidence." *Philosophy of Science* 67 (Proceedings): S193-S207.

Morrison, Margaret. 2007. "Where *Have* All the *Theories* Gone?" *Philosophy of Science* 74 (2):195-228

Noë, Alva. 2005. *Action in Perception*. Cambridge, MA: MIT Press.

Peschard, Isabelle. 2011. "Modeling and Experimenting." In *Models, Simulations and Representation,* ed. Paul Humphreys and Cyrille Imbert, 42-61. London: Routledge.

Rooney, Phyllis. 1992. "On Values in Science: Is the Epistemic/Non-Epistemic Distinction Useful?" in *Proceedings of the 1992 Biennial Meeting of the Philosophy of Science Association*, ed. David Hull, Micky Forbes, and Kathleen Okruhlik, 13-22. East Lansing, MI: Philosophy of Science Association.

Staley, Kent. 2005. "Agency and Objectivity in the Search for the Top Quark." In *Scientific Evidence: Philosophical theories and applications,* ed. Peter Achinstein, 165-184. Baltimore: The Johns Hopkins University Press.

Suppes, Patrick. 1962. "Models of Data." In *Logic, Methodology, and Philosophy of Science*, ed. Ernest Nagel, Patrick Suppes, and Alfred Tarsi, 252-61. Stanford: Stanford University Press.

Varela, Francisco, Evan Thompson and Eleonor Rosch. 1991. *The Embodied Mind.* Cambridge, MA: MIT Press.

Wheeler, Gregory. 2000. "Error Statistics and Duhem's Problem." *Philosophy of Science*, 67(3): 410–420.