Evidential Diversity and the Triangulation of Phenomena

Jaakko Kuorikoski and Caterina Marchionni*†

The article argues for the epistemic rationale of triangulation, namely, the use of multiple and independent sources of evidence. It claims that triangulation is to be understood as causal reasoning from data to phenomenon, and it rationalizes its epistemic value in terms of controlling for likely errors and biases of particular data-generating procedures. This perspective is employed to address objections against triangulation concerning the fallibility and scope of the inference, as well as problems of independence, incomparability, and discordance of evidence. The debate on the existence of social preferences is used as an illustrative case.

1. Introduction. Approaching the phenomenon of interest from diverse perspectives and using multiple methods to investigate it is often referred to as *triangulation*. Although it is a term routinely used by scientists (see Blaikie 1991; Hammersley 2008) and sometimes also by philosophers (see Wimsatt 1981; Schickore and Coko 2013), there is no agreed-on definition of what it amounts to or a shared understanding of the basis of its epistemic value. The idea of the epistemic virtue of using diverse evidence to increase the con-

Received March 2015; revised September 2015.

*To contact the authors, please write to: Jaakko Kuorikoski, PO Box 24, University of Helsinki; e-mail: jaakko.kuorikoski@helsinki.fi. The names are in alphabetical order, and the authors contributed equally to the article.

†This article was presented at the Causality and Experimentation in the Sciences conference in Paris 2013, the CamPos seminar in Cambridge 2014, the British Society for Philosophy of Science conference in Cambridge 2014, the Evidence in Science and Epistemology workshop in Helsinki 2014, and the Working Seminar on Robustness Analysis in Helsinki 2014. We thank the audiences in these meetings for their comments. We would especially like to thank Casey Helgeson, Chiara Lisciandra, Aki Lehtinen, Jonah Schupbach, Kent Staley, Jacob Stegenga, and two anonymous referees for their invaluable constructive critiques. This research has been supported by the Academy of Finland. Caterina carried out part of this work while visiting the Centre for Philosophy of Natural and Social Science, London School of Economics.

Philosophy of Science, 83 (April 2016) pp. 227–247. 0031-8248/2016/8302-0004$10.00

firmation of hypotheses is intuitively plausible—even taken as a trivial truism by many—and yet the concept remains ambiguous and contested. In methodological discussions in the social sciences and recently also in the philosophy of science, formidable criticism has been leveled against the usability of triangulation. Furthermore, those who agree on the epistemic importance of diversity of evidence disagree about what exactly this value is based on. In this article, we propose an account of the epistemic rationale of triangulation as a form of robustness reasoning. That is, we understand triangulation as the use of multiple and independent sources of evidence to ascertain whether a phenomenon is an artifact of a particular method of determination (Campbell and Fiske 1959; Wimsatt 1981). This notion of triangulation is closely related to that of evidential diversity, but although the confirmational significance of evidential diversity is a widely accepted epistemic principle (e.g., Fitelson 2001), several worries about triangulation have been voiced. For example, Stegenga (2009) claims that triangulation faces several difficulties that limit its epistemic value: evidence produced with different methods is often incomparable, a criterion of independence is needed but is not available, triangulation does not always work as a confirmatory procedure, and multiple methods often yield results that are incongruent. Hudson (1999, 2009, 2013) claims that triangulation has no epistemic value and that despite occasional allegations to the contrary, scientists do not actually make use of it.

We address these and other worries and show that they arise from either too narrow or too broad a conception of the methodology of triangulation. We propose that in order to adequately assess the epistemic import of triangulation, two types of evidential reasoning need to be distinguished: data-to-phenomenon inferences and phenomenon-to-theory inferences (Bogen and Woodward 1988). In the standard fashion, by 'phenomena' is here meant robust and (at least partially) replicable results, effects, or regularities. Phenomena are (usually) not directly observable but must be inferred from data. Theories explain and are tested by phenomena, and the existence of phenomena is inferred from data. The two types of inference are different: the first type of inference concerns the deductive (or broadly logical) relationship between a representational system (a theory) and a description of a phenomenon, whereas the latter concerns the reliability of the causal processes by which the data are generated.

We argue that triangulation is to be understood as a methodological strategy employed for the purpose of controlling for errors and biases of particular methods of determination. The notion of error control here is epistemic, not concrete: independent experimental procedures or kinds of evidence can be used to increase the 'aggregate' reliability of data-to-phenomenon inferences. These inferences concern the causal processes generating the evidence. In the triangulation of phenomena, what we need to worry about are *errors and biases* in the particular processes at play. Accordingly, the relevant notion of

independence is that of (causal) *reliability independence*, which does not require knowledge of all the background assumptions, as is (arguably) required for the confirmational boost that explaining or predicting a variety of phenomena confers to a theory. Reliability independence is established on a case-by-case basis and does not pose any general in-principle issue of incomparability of evidence. Furthermore, in data-to-phenomenon inferences, the confirmational added value of triangulation can be captured by the probabilistic condition of confirmational independence, closely related to the causal screening-off condition. We illustrate our claims with an example drawn from the behavioral sciences: the use of behavioral, psychological, and neurological data as allegedly diverse evidence in support of the existence of social preferences.

2. Triangulation as a Robustness Argument. In broad terms, triangulation can be understood as the use of multiple perspectives to gain a more comprehensive understanding of something. Here we endorse Campbell and Fiske's (1959) narrower definition of triangulation as *the use of multiple methods to measure or detect the same property or phenomenon*. So defined, triangulation is a form of robustness analysis: although each of the multiple methods is a source of possible error, if the same inference is drawn regardless of which error liable method is used, then the errors of the methods used probably do not matter to the inference at stake. This sense of triangulation is to be distinguished from other ways in which diverse evidence may be epistemically useful, such as providing novel information about the phenomenon under investigation and thus increasing overall understanding, providing evidence that some particular method is reliable (see sec. 4.6), consilience of inductions or unification in the sense that a hypothesis explaining a diverse set of phenomena by itself increases its confirmation, and determining that the phenomenon itself is causally robust under diverse disturbances and background conditions. Our interest is only in the epistemic added value that diverse and independent sources of evidence provide for a single hypothesis (about the same property of the phenomenon under investigation).

In order to properly locate our analysandum, it is worth making a few initial observations concerning robustness analysis in general. First, robustness analysis has recently been the focus of sustained philosophical attention, but different bodies of literature refer to rather different practices under the general category of robustness analysis (e.g., Wimsatt 1981; Weisberg 2006; Woodward 2006; Stegenga 2009; Soler et al. 2012; Hey 2015).[1] According

1. For example, Woodward (2006) distinguishes five types of robustness analysis. His notion of *measurement robustness* is the one that comes the closest to our characterization of triangulation, but whereas the former focuses on measurement procedures (namely, assigning numerical values to quantities), we are interested in the establishment of the

to some, a common rationale undergirds various forms of robustness analysis (e.g., Wimsatt 1981; Schupbach 2015). However, others argue that inferences that are legitimate on the basis of one kind of robustness may not be so in terms of another. Our view is that even if various forms of robustness analysis did share a common broad rationale, identifying such a rationale would not as such suffice to yield the conditions under which robustness-based scientific inferences are legitimate. In other words, there are two distinct philosophical projects: one is to identify at an abstract level the logic behind successful robustness arguments, while the other is to determine what is required for a specific form of robustness analysis to be successful. This article is a contribution to the latter project.

Second, critics of robustness analysis sometimes argue as if to establish that robustness analysis has little or no epistemic value, it is sufficient to demonstrate that it can fail to deliver the truth. However, robustness analysis should not be regarded as a procedure for conclusively establishing the truth of a result, a hypothesis, or a theory but rather one aimed at increasing the reliability of inferences from less than perfectly reliable scientific tools. Consequently, triangulation is a heuristic that should be expected to work well in a certain class of cases yet is neither necessary nor sufficient for establishing truth (Wimsatt 2007). Thus, the challenge is to identify the circumstances in which confidence in the result rationally increases on the basis of the concordance between independent methods of determination.[2]

Third, robustness analysis has been thought to be an instance of a no-miracles argument (e.g., Hudson 2013). This association has been possibly inspired by the use of the robust determination of Avogadro's constant as an argument for the molecular theory of matter as a paradigm case of triangulation reasoning. We argue, however, that this is misleading. First, interpreting the 'no-miracles argument' as an inference to the best explanation is subject to the quite obvious objection that if all the methods of determination are wrong, the explanation of their agreement might be due not to the truth of the result but to their sharing a common mistaken assumption (see also Salmon 1984; Cartwright 1991). Second, the very idea of a substantial inference to the best explanation is highly suspect. In order to identify when confidence in a conclusion is justifiably increased, we have to go beyond the intuitive no-miracles argument. This is what we endeavor to do in the following sections.

---

reality of (mainly experimental) phenomena more generally. Likewise, Schickore and Coko (2013) use the term *triangulation* to refer to "the use in empirical practice of multiple means of investigation to validate an experimental outcome" (296).

2. We are not committed to a fully Bayesian view of scientific reasoning and of the aims of inquiry, however, but only to the broad idea that it makes sense to discuss comparative weighing of evidence. The latter is a necessary condition for triangulation and related heuristics.

3. Evidential Diversity and Data-to-Phenomenon Inferences. The first building block of our account of triangulation is the confirmational independence of diverse evidence. One can formulate this independence condition probabilistically:

> E1 and E2 are confirmationally independent with respect to a hypothesis H if and only if the amount of confirmation (given one's choice of measure) provided to H by E1 is independent of E2 and vice versa ($c(H, E1|E2) = c(H, E1)$ and $c(H, E2|E1) = c(H, E2)$). A sufficient condition for confirmational independence is that E1 and E2 are probabilistically independent conditional on H (or not H) ($E1 \perp E2 \mid H$), but dependent unconditionally (i.e., H screens off E1 and E2; Sober 1989; Fitelson 2001).

This means that, given that the result is known (in the binary case, that the hypothesis is true or false), whether one means of providing evidence produces the 'right' result does not affect the probability of the other getting it right. If this holds, the confirmation provided by the different methods is independent of each other. Such multiple independent pieces of evidence provide additional confirmation to the hypothesis, over and above the individual pieces of evidence. This is a relatively robust principle in line with what we take to be plausible candidates for theories of *degree* of evidential support (such as Bayesianism, with likelihood ratio as the measure of confirmational support, and likelihoodism).[3] Rather than diversity per se, it is this independence condition that grounds the confirmatory added value of a variety of evidence (Fitelson 2001; see, however, Bovens and Hartmann 2002).

As such, however, the formal confirmational independence condition does not determine what the probabilities are attached to and, consequently, how the independence is and should be realized in practice. The second building block of our account of triangulation addresses this issue. We argue that insofar as triangulation pertains to data-to-phenomenon inferences, which are essentially causal inferences, then there is a rather intuitive way to see how independence can be realized, one that, as we will show, blocks many of the criticisms leveled against triangulation as a robustness argument.

In the vein of Bogen and Woodward's account, data-to-phenomena inferences are causal inferences in the sense that they involve empirical reasoning about the particular causal processes that generate the data and their possible errors as well as the experimental manipulation of the phenomenon of interest (Bogen and Woodward 1988; Woodward 2000). Hence, the evidential rele-

3. A similar principle can also be rationalized in the severe testing framework, but strictly speaking, any discussion of degree of support would then be out of place. Staley (2004) rationalizes the value of triangulation within a severe testing framework. We discuss the relationship of Staley's and our accounts in sec. 4.6.

vance of data to phenomena cannot be informatively reconstructed solely as a set of entailment relations between theoretical and observational sentences. The establishment of deductive relations between phenomena and data claims is at least not necessary for the evidential relevance of the latter: scientific practice shows that the evidential status of data does not require full theoretical knowledge of the principles of the data-generating processes.[4]

Next, we argue that if triangulation pertains to data-to-phenomenon inferences, the required screening-off condition, and consequent confirmational independence, is satisfied when the different methods producing the data (say, E1 and E2) do not share any systematic errors and biases. This means that, with respect to causal inferences from data to phenomena, confirmational independence is realized by what Schupbach (2015) calls *reliability independence* (i.e., that the characteristic biases and errors of the different methods are independent). To see this, consider (a property of) a phenomenon investigated by two methods. The methods are used to produce data (evidence) about the phenomenon such that systematic patterns in the data track the property of the phenomenon. The property is therefore a common cause of the two types of evidence. Let us think of this in terms of random variables forming a conjunctive fork.[5] Since the methods are not fully reliable, this conjunctive fork is noisy—one can think of this in the standard way as additional error terms added to the evidence variables. The phenomenon variable screens off the evidence variables if these error terms are independent. We can now distinguish two reasons why such independence should be expected to hold. First, if the processes of data production (methods) are causally independent (being based on different kinds of causal mechanisms), then any token random causal disturbance of one method should not have an effect on the other method. Second, if the methods are based on different kinds of causal processes, the presence of any systematic error (bias) in one method should not affect the probability of an error occurring in the other. Thus, if the methods are not prone to similar disturbances (errors) and systematic biases, the error terms are independent of each other, in which case the conjunctive fork realizes the confirmational independence condition. Thus, confirmational independence is realized by reliability independence.

How do scientists know whether the methods or instruments share the same error, so that their agreement constitutes a reliable basis for triangulation? In causal data-to-phenomenon inferences, the epistemic task is to distinguish signal from noise and to control for biases and errors in the causal

4. We take causal reasoning to be essentially characterized by the interventionist theory (Pearl 2000; Woodward 2003), although none of the central steps in our argument depend on the details of that theory.

5. Of course, this means accepting the principle of common cause.

production of the data. Such causal inferences take into account the process that produced the data, which involves a number of context-dependent considerations (Bogen and Woodward 1988; Woodward 2000).[6] Likewise, triangulation should take into account context-dependent considerations about the particular processes that produce the data and their common (as well as different) sources of possible errors. Moreover, the fact that triangulation concerns a comparison between particular causal processes means that two or more processes of data production being independent in the relevant way depends on what we are trying to detect or measure. Thus, two (or more) methods may be independent in providing evidence for P but not for P*. For example, different brain-imaging techniques (such as fMRI and MEG) can be relatively independent with respect to measurements of brain area activation because they rely on different types of physical processes (fMRI tracks blood oxygenation, whereas MEG directly tracks the minute currents produced in neuronal communication) but dependent when taken as methods of determining cognitive function (because as means of detecting specific cognitive functions, they rely on the same localization hypotheses).

Second, agreement of results is necessary for successful triangulation, but it is not sufficient. It is independence with respect to types of error that matters. The identification of errors occurs via local empirical investigation, that is, by means of replication, calibration, and similar strategies well known to scientists (see Woodward 2000). There are cases in which scientists have enough knowledge of their experimental and measuring apparatuses to be quite confident about the independence of errors that need to be controlled for. In these cases, the inference is quite strong. In other cases, more empirical investigation is required regarding the probability of errors that the particular experimental apparatuses may encounter in relation to the detection of the property or phenomenon at hand. The more empirical (experimental) or theoretical reasons there are to suspect that the different methods might share relevant errors and biases, the less confidence one should have that the convergence of results is due to the phenomenon rather than the shared errors. This also means that whether one method produces the result will have a bearing on the probability that the other method produces the result, given knowledge of the phenomena they are both measuring. Consequently, confirmation provided by one method is also less independent of the confirmation provided by the other. The epistemic added value of triangulation comes in degrees, and full reliability, and hence confirmational, independence should be taken as the ideal limiting case.

6. Woodward (2000) characterizes them as 'empirical' inferences in contrast to the 'logical' inferences that are emphasized the most in the context of theory confirmation, although of course not all methods of theory confirmation are based on logical inference alone.

It is only once a phenomenon is inferred to be robust (i.e., a 'real phenomenon' and not an artifact) that our theories are expected to be able to account for it (see Campbell and Fiske 1959; Bogen and Woodward 1988). Whatever one's preferred method of confirmation and theory testing, it can be argued that 'accounting for the phenomenon' provides confirmation to a theory. If it is hypothetico-deductivism, then the relevant relationship is one of deductive implication. If a Bayesian theory of confirmation, then it is a matter of appropriate subjective probabilistic conditional dependencies (Jeffrey conditionalization could be used to show how the probability of a theory increases when the probability of a phenomenon claim is increased via triangulation).[7] Something like the logic of severe testing is also applicable: theories are best confirmed by the kinds of phenomena that would be (given background knowledge and rival theories) very unlikely to exist if the theory were false. We need not commit to any of these views about theory confirmation here, however. What matters for our purposes is that the relationship between theory and phenomenon is not a causal one and, therefore, that the ideas of error control and the independence of phenomena have to be interpreted differently.

4. Six Objections to Triangulation. We have laid out the bare bones of an account of triangulation as causal reasoning from data to phenomenon, the epistemic rationale of which lies in controlling for likely errors and biases of particular data-generating procedures. In the limiting case in which different methods are based on completely independent causal processes, the methods do not share any errors and biases and are, as a result, confirmationally independent. We now show that on this understanding of triangulation, many of the worries that have been voiced against it turn out to be unfounded.

4.1. Two Wrongs Do Not Make a Right. The first line of skepticism against the value of triangulation claims that the very possibility that multiple incorrect methods may nonetheless produce a congruent result speaks against the epistemic value of triangulation (e.g., Cartwright 1991; Hudson 1999, 2009, 2013). In an elaboration of this criticism, Hudson (2013) attempts to show that historical cases of robustness reasoning can be better rationalized as aimed at checking the reliability of a primary method by other methods, a procedure Hudson considers as 'calibration'. The general argument is that if one method is reliable, then we do not need other methods, but if no method is reliable, then we do not gain anything by adding more bad methods on top of existing bad ones. In our account, triangulation is a matter of *degree of support*. As long as the methods are at least approximately in-

7. We thank an anonymous referee for suggesting this.

dependent, the claims that multiple (less than perfect) methods do not possess any value and that adding methods alongside a reliable one does not increase overall reliability are simply false.

4.2. Independence. A second source of skepticism focuses on the very possibility of individuating methods of data generation and establishing their independence. Stegenga (2009, 2012) proposes defining distinctness and independence in terms of problematic-auxiliary assumptions: two methods of determination are distinct and independent when their problematic auxiliary assumptions are independent relative to a given hypothesis: "A hypothesis is more likely to be true when two or more modes of evidence provide concordant multimodal evidence and the worrisome or problematic assumptions for all modes of evidence are independent of each other" (Stegenga 2012, 219). We assume that by 'independence' Stegenga means that the assumptions can be true or false independently of each other. Although this is in line with the general epistemic rationale of the variety of evidence principle, Stegenga finds it problematic that in applying such robustness arguments we do not know how to identify the problematic assumptions. Therefore, what he calls the 'problem of individuation' is shifted back from having a general criterion for identifying methods to having one for identifying problematic background assumptions.

Stegenga's requirement, however, is unnecessarily demanding, both in requiring that all background assumptions be controlled for and that a general criterion of identification exists. It is unclear why we need to know that all the assumptions underlying each method of determination are independent (see also Claveau 2011, 243). If confidence in a result is a matter of degree, then it ought to depend on the number and, more importantly, the types of assumptions that are independent. As we have seen above, in data-to-phenomenon inferences, learning about independent reliability is an empirical matter. Of course, in most cases not all sources of error are known and controlled for. But does this undermine the epistemic value of triangulation? It does not. Even if we take near certainty to require the triangulation of a result through as many methods as there are problematic assumptions or possible sources of errors, incremental increases can be achieved with less than this (see Kuorikoski, Lehtinen, and Marchionni 2010, 2012; Odenbaugh and Alexandrova 2011).

Second, why should there be a general criterion for identifying problematic assumptions? In principle, it is possible to learn empirically about whether particular processes of data production are independent and whether they share the same sources of systematic biases and errors (e.g., by calibrating them to an already established phenomenon or by manipulating other factors known to affect the phenomenon and checking whether the data-generating processes produce suitable systematic variation). These empirical

investigations do not provide us with a general criterion for identifying problematic assumptions but contextual knowledge about which assumptions are more likely to matter. In practice, theoretical arguments for and against independence of methods are usually presented, but in many cases there is only incomplete, or even completely wrongheaded, theoretical understanding about why the detection methods work as they do (see, e.g., Hacking's discussion of the history of the microscope [1983, chap. 11]). For example, the reliability of what has now become the true workhorse of cognitive neuroscience, fMRI, was long held to be suspect because of incomplete theoretical understanding of the physical basis of the BOLD signal and its detection. What finally convinced the brain research community of the reliability of the method was not any breakthrough in the theoretical understanding of the underlying physics but the experimental establishment of a reliable connection between the direct electrical stimulation of a macaque brain and the resulting fMRI images. The important point is that the very nature of the problem of identifying background assumptions changes when triangulation is investigated in the context of causal reasoning from data to phenomenon.

4.3. The Screening-Off Condition. Hudson offers the following reductio argument against the probabilistic confirmational independence condition that we favor. If knowing the true value of the property being determined by the independent methods screens off the probabilities of the outputs of the methods, then knowing the true value ought to screen off one measurement produced by one method from another produced by the very same method. In other words, if it is independence rather than variety that is responsible for the increase in confirmation (as we claim), then one should be able to triangulate with only one method, provided that the outputs are suitably independent. And triangulating with one method is surely nonsensical (Hudson 2013, 18–20).

We agree that if the confirmational independence condition holds, one can triangulate with only one method. But that is a big *if*. The condition in effect demands that the information conveyed by the different observations is, well, independent. The only way that successive outputs of a single method can satisfy this condition is that every output provides genuinely new information about the target. In our account, the purely formal confirmational independence condition is fulfilled when the different methods qua causal processes do not share the same errors and biases: whether one method produces a 'faithful' mark of the target is independent of the success of the other. If successive applications of one and the same method were independent in this way, either the method would actually be free of systematic errors (unbiased) or each application of the method would somehow constitute a different kind of causal process sharing no biases with the other applications. In the first case, simply repeating the use of the same method would in fact be the epi-

stemically rational thing to do. For example, consider the simple case of determining a population statistic, such as the mean: genuinely independent and unbiased samples are the best means of determining the statistic, and adding variety to the methods of determination would only be counterproductive. In the second case (i.e., each application would constitute a different kind of causal process sharing no biases with the other applications), it is hard to see how the different observations could be said to have been produced by the same method.

Schupbach (2015) also argues against the confirmational independence condition. Unlike Hudson, who challenges the sufficiency of the screening-off condition, Schupbach questions its plausibility in practice: the results of different methods are rarely (if ever) going to be fully independent conditional on the negation of the hypothesis, since there are always other possible reasons for the results to be dependent. Schupbach considers the case of the observation of Brownian motion. He points out that observing motion in different types of pollen does not satisfy the confirmational independence condition, since the common result (spontaneous motion of the particles) could be plausibly caused by some other shared features of the methods, such as a vital force inherent in the pollen or some artifact of the shared suspension medium. Since learning about the reliability of the experiment with one type of pollen is relevant to the assessment of the reliability of experiments employing other types of pollen, the results are not independent conditional on the negation of the Brownian motion hypothesis. Hence, Schupbach's argument goes, in this paradigmatic case of robustness reasoning (and in others such as the Lotka-Volterra models) the confirmational independence condition does not capture the relevant kind of independence. In turn, this casts doubts on whether confirmational independence is the right descriptive account of robustness reasoning in science.

Schupbach makes an important point here, but the problem is not as severe as he suggests in the case of causal inferences about phenomena. Data are evidence of a phenomenon if variation in the data systematically tracks variation in some aspect of the phenomenon due (only) to the causal process that generates the data. Therefore, the negation of the hypothesis does not include the whole set of logical possibilities consistent with the phenomenon not being present. The results only need to be independent with respect to a set of plausible common causes (confounders) capable of producing the kind of variation that could be mistaken to be caused by the phenomenon.[8] In the pollen example, surely the suspected vital force could create the kind of variation that could be mistaken for the effect of random molecular col-

8. As an extreme example, the presence of oxygen would be tracked by any two experiments because of the scientists' need of oxygen. Nevertheless, oxygen in this trivial sense is not usually taken as a serious confounder in experimental practice.

lision, but varying only the type of pollen implies that only the possible errors related to the specific kind of particle are being controlled for. Therefore, although varying the type of pollen certainly is an example of robustness reasoning, it is not a paradigmatic case of triangulation, which aims at independence with respect to as many relevant causal features of the different data-generating processes as possible as a means to control for all possible (esp. yet-unknown) errors. The whole arsenal of types of observation and experiment used by Perrin in determining Avogadro's number better exemplifies the logic of triangulation.

According to Schupbach's positive proposal, the rationale of robustness analysis lies in the efficiency with which independent methods eliminate salient alternative explanations of the observed phenomena. In the previous example, the different pollens are 'eliminatively diverse' in that they rule out the possibility that the result is due to the form of a particular type of pollen. Although full reliability independence is not satisfied in this case, there are nevertheless important similarities between our and Schupbach's accounts. First, according to our account of triangulation, the observed phenomena are the congruent data produced by the different methods, and the salient competing explanations for the observed results are precisely the characteristic errors and biases of particular methods. Second, when full confirmational independence is satisfied, all alternative causes of the convergence of the results are in fact ruled out.[9] Hence, triangulation also works by eliminating the biases and errors of particular methods from the list of salient causes of the observed results. Third, in cases of failure of robustness, the kind of eliminative reasoning exemplified by varying only a particular aspect of the data-generating process (e.g., the type of pollen) is more informative since it allows scientists to locate the source of the error. Schupbach's eliminativist account is therefore not a competing alternative: both piecemeal eliminative robustness analysis and triangulation are forms of robustness reasoning, but they are still different epistemic strategies.

4.4. Incomparability. The incompatibility of ontologies is a common argument against the feasibility of triangulation in the social sciences (see Blaikie 1991). Stegenga (2012) points to a similar problem: if the data are produced with methods based on completely different kinds of causal processes and relying on different (and possibly mutually inconsistent) back-

9. In our view, Schupbach's eliminativist rationale for robustness analysis is in fact at least compatible with the probabilistic rationale for the epistemic added value of diversity provided by confirmational independence. As Fitelson puts it, the core idea captured by the condition is that the advantage of diversity arises from "data whose confirmational power is maximal, given the evidence we already have" (2001, 131). Whether the condition quantitatively captures the very same epistemic gain as eliminative diversity has to be left for further research.

ground assumptions, then how can the data be about the same phenomenon? Getting different kinds of data to speak meaningfully with each other is usually not a trivial task, and some skepticism toward triangulation is warranted when insufficient care has been taken to ensure that interesting inferences can be justifiably drawn from collating diverse data. In our understanding of triangulation, however, there need not be any tight conceptual 'mapping', such as deducibility, or some kind of theoretical reduction between the ontologies of the phenomenon and the data. What is needed is only (preferably controlled) covariation between the different kinds of evidence and plausible background theories concerning the causal processes between the phenomenon and the data. Again, the experimental demonstration of the ability of fMRI to track neural activity shows that experimental control of the target system can compensate for the lack of theoretical mapping between the different methods and the target phenomenon: wiggle the target (possibly a more easily accessible surrogate system, like the monkey brain), and if the different methods produce correlating outcomes, we have a (defeasible) reason to believe that they track the same property.

4.5. Discordance. Even though by definition triangulation requires the results of different methods to be congruent, sometimes there are results that seem to point elsewhere than the majority. So is triangulation futile whenever there is an indication of discordance among the plethora of methods? According to Stegenga (2009, 2012), when different methods produce discordant rather than concordant evidence, it seems that there would have to be some way of comparatively evaluating the weight of individual pieces of evidence for triangulation to rationally increase our confidence about the hypothesis. In turn, this would require us to know the 'amalgamation function' for the separate pieces of evidence, and it is quite obvious that we are rarely in possession of such a function. Without a rational way of comparatively weighing evidence, so the argument goes, appeals to triangulation amount to no more than hand-waving. However, even if, as seems likely, cases of discordant evidence occur more frequently than cases of concordant evidence, discordance does not pose an insurmountable problem for triangulation.

In principle, the epistemic rationale of triangulation works just as well when some of the evidence independently lowers the probability of the hypothesis; the problem is evaluating the extent to which the discordant evidence matters. On one hand, Stegenga is right in that we are rarely in a position to provide a full amalgamation function for all of the evidence, so the lesson of discordance is that in such cases, we should only be as confident in the implications of the robustness of the result as we are confident about the relative differences we assign to the reliability and weight of the different evidence. On the other hand, if we have no principled way of making such comparisons, we should not be talking about increasing or decreasing degrees

of support in the first place. In such cases, the only rational thing to conclude from discordance is that something is amiss and hence that it is rational to suspend any judgment on the hypothesis—at least for the time being (Hey 2015). But if we cannot assign any differences to evidential weight to begin with—and implicit equal weighing is weighing nonetheless—triangulation does not make any sense, regardless of discordance.

4.6. Security and Reliability. The main competing rationalizations for what at the outset seems like triangulation appeal to the use of independent methods to increase the reliability of a primary method, which, in the end, is the one providing the evidence for the hypothesis. According to Staley (2004), rather than directly increasing the evidential support ('strength') of a hypothesis, the main role of triangulation is to increase the 'security' of a primary evidential claim, that is, "the degree to which [a] claim [about particular data being evidence for a hypothesis] is susceptible to defeat from the failure of an auxiliary assumption" (468).[10] Hudson (2013) also argues that in the historical episodes that philosophers (and scientists, for that matter) have identified as instances of robustness reasoning, including Perrin's much celebrated determination of Avogadro's number, robustness reasoning was not in fact involved and that the function of multiple ways of measurement was just to test the reliability of the primary method. Is there a principled way of keeping triangulation apart from tests of procedures, or do the two confirmatory effects—the one for the reliability of methods and the one for the phenomenon—amount to the same thing after all?

According to Staley (2004), convergent results can increase the security of an evidential claim in two ways. First, results from an alternative, independent test can be used as evidence for one or more assumptions behind the primary evidential claim. Second, the result from a second test can function as 'backup' evidence in case one or more assumptions of the first test turn out to be false (474). Although our account addresses the direct evidential support provided by congruent yet independent results, it is clearly compatible with the backup route, since useful backup evidence should be confirmationally independent—the main difference being that we are not committed to any epistemically relevant ordering of methods into primary and secondary. Furthermore, our account does not rule out the possible relevance of an additional congruent result for increasing the security of evidential claims based on other methods.[11] But when do congruent results provide direct support and when do they increase security?

10. The failure of an auxiliary assumption falls within our notion of error.

11. Note that this is not inconsistent with the confirmational independence condition, since the condition is defined in the (idealized) situation in which the hypothesis (phenomenon claim) is known to be true or false.

Whether concordant results from independent tests confirm the phenomenon hypothesis or increase our faith in the reliability of the methods depends on (i) the prior credibility of the result and (ii) the prior assessment of the reliability of the methods. The higher scientists' knowledge of the methods and their reliability, the more their convergence confirms the result and vice versa. If both the result and the methods are regarded to be highly unreliable, possibly for similar reasons, scientists have no legitimate grounds to take convergence to be evidence for the result. There is no reason to assume that scientists ought to be testing either the reliability of inferences about phenomena or the reliability of the procedures and not both (Franklin 1986). However, this does not mean that the scientist chooses how to use the results at hand: congruence of results does not come as a pie of confirmation that the scientist can apportion to alternative epistemic aims as she pleases. Moreover, the congruence of results does not warrant a kind of bootstrapping inference, that is, the use of the congruent results as grounds for inferring the reliability of a method in order to further increase the credibility of the results.

We do not want, nor do we have space here, to engage Hudson's historical interpretation of Perrin's argument as a case of 'calibration'. We do, however, want to comment on Hudson's argument against the plausibility of triangulation in general. Hudson's main argument is the well-known one that if none of the procedures by which the result is obtained are reliable, then convergence of their results does not have any epistemic significance (Hudson 2013; see also Cartwright 1991). In response, we simply reiterate the point that we already made: this is simply false if we can reasonably entertain degrees of support and if the confirmational independence condition holds (at least approximately). If triangulation is understood, as we do, as a matter of controlling for the likely errors of diverse procedures, then it is possible to identify the conditions under which the inferences are stronger. In particular, as we have argued above, the more certain scientists are that the procedures do not share the same kind of error or bias and that each procedure is itself reliable to a degree, the stronger the inference to the existence of a phenomenon will be.

5. Triangulation in Action: Neuroeconomics of Social Preferences. Our discussion has so far remained at an abstract level. We now illustrate some of our claims in the context of an ongoing debate involving economics, psychology, anthropology, and neuroscience: the existence and causal role of social preferences. Social preferences refer, albeit somewhat ambiguously, to overt behavior, as well as sometimes to its underlying motivational states, such that the agent takes into account (either positively or negatively) the material payoff or the well-being of others (Fehr and Krajbich 2014, 193). Such states are postulated to explain ubiquitous deviations from standard microeconomic predictions of individual choice behavior in the laboratory, as

well as in the field. People donate to charity, share when they do not have to, and punish those who act unjustly even when this is costly to themselves. What exactly people are thought to care about varies from the well-being of others (altruism) to overall fairness (inequity aversion). A set of standardized game-theoretical experiments, such as the dictator game and the ultimatum game, are interpreted to measure the strength of these motivations across a wide range of cultural contexts.[12] Although the idea of social preferences is hardly counterintuitive, the stability and the very existence of these preferences are still widely questioned.

Rival explanations for apparently pro-social behavior in the laboratory include importation of mistaken social scripts and framings to one-off strategic interaction (Binmore 2005), pursuit of social esteem (Andreoni and Bernheim 2009), and the motivation to play the game 'correctly' in the very artificial experimental situation (Levitt and List 2007). In other words, these alternative explanations point to the possibility that pro-social behavior is an artifact of the experimental setup. In the case of observed pro-social behavior in the wild, the alternative explanatory hypotheses are underdetermined by the observational data. Furthermore, the within-subject stability of these preferences across different games is relatively poor, which casts doubt on the experiments' ability to measure stable motivational features that could be taken to explain pro-social behavior in the wild.

Accordingly, even though social preferences are very well documented when interpreted purely behaviorally, their interpretation as motivations is more contentious. In particular, the controversy is about whether they are preferences to begin with (i.e., part of the goal-directed deliberate repertoire of behaviors) and, if they are preferences, whether they are truly other-regarding rather than fundamentally selfish.

In a recent review of the neuroscientific literature on social preferences, Fehr and Krajbich claim that "one emerging theme of the studies reviewed is that social reward activates circuitry that overlaps, to a surprising degree, with circuitry that anticipates and represents other types of rewards. These studies reinforce the idea that social preferences for donating money, rejecting unfair offers, reciprocating others gifts, and punishing those who violate norms are genuine expressions of preferences" (2014, 214). The idea is that by identifying the neural circuitry that is activated when individuals display pro-social behavior, insights can be obtained into whether their motivations are genuine expressions of social preferences. In this sense, neuroeconomics

12. In a dictator game, the first player (proposer) simply decides whether and how much to give to the other player (responder). Rational choice theory predicts, assuming self-interest, that the proposer gives nothing and keeps all of the endowment. In an ultimatum game, the responder has the option of denying the offer, in which case neither player gets any money.

experiments are regarded to contribute to purely behavioral experiments not only in terms of providing mechanistic explanations (Craver and Alexandrova 2008; Kuorikoski and Ylikoski 2010) but also by providing additional evidence that pro-social behavior is not an experimental artifact to begin with. Woodward (2009) argues that neural evidence can contribute to the triangulation of social preferences, in particular regarding motives and preferences. For example, neural evidence might provide information about the motivations underlying low rejections in ultimatum games: whether they involve inequality aversion or a taste for negative reciprocity (198).

However, the use of neuroeconomics evidence for the triangulation of social preferences faces three challenges. First, for successful triangulation we need independent evidence, but is neural evidence on social preferences suitably independent from behavioral evidence? Ross's (2008) complaint that many neuroeconomics experiments simply replicate well-known experimental games, with the only difference being that the subjects' brain activity is scanned, would seem to suggest that it is not. Second, it has been argued that there are conceptual issues that make the neural evidence about behavior hard to compare with more traditional economic evidence (e.g., Fumagalli 2013, 336). Third, much of the debate around neuroeconomics has concerned its relevance to economic models and theories, which are about observable choices and make no prediction about the brain (e.g., Gul and Pesendorfer 2008; however, see Clarke 2014). Therefore, it is unclear whether neural data can be used as evidence for or against economic theories or models.

How does our account help to resolve these problems? Starting with the last one, since triangulation pertains to inferences from data to phenomenon, we do not need to worry about whether neural data can provide direct confirmation to economic theories and models. This is not to say that this is not an interesting problem but only that our account of triangulation is silent on this question.

The second issue pertains to the comparability of different kinds of evidence. For example, Fumagalli (2013, 336) notes that "profound conceptual differences remain between a variable such as the firing rate of some neurons and an abstract construct like decision utility (e.g. only the former is a cardinal object)." But triangulation as we characterized it does not require a tight conceptual mapping from the ontology of the evidence statements (e.g., about brain activation) to the ontology of the theory to be tested (e.g., decision utility). This is because we conceive of triangulation as a species of causal reasoning from data to a phenomenon, and what therefore suffices are the empirical principles of causal reasoning, especially when it comes to reliability and error control. Hence, the issue of comparability turns into the task of establishing whether diverse pieces of evidence are generated by the intended common cause. For example, imaging data tracking such differences in affect-related areas (anterior insula or anterior cingulate cortex) that

correspond to differences in pro-social behavior (e.g., whether the subject engages in costly punishment of antisocial behavior) are, or so the hypothesis goes, caused by the affective states that partly constitute the social preference for fairness. More generally, as was stated above, the credibility of basic imaging techniques (e.g., fMRI) as a means of producing data relevant for cognitive and affective functions was not conclusively established through a theoretical argument but by the experimental manipulation of a surrogate system (a direct stimulation of a macaque brain) and a corresponding difference in the fMRI output.

The remaining problem concerns the independence of neuroeconomics experiments vis-à-vis behavioral experiments. Obviously neuroeconomics results based on the same kind of experimental games as the behavioral experiments do not contribute much to claims about the pro-social behavior in the lab. The crucial question then is whether they can be independent in a way that increases the reliability of inferences about what drives such behavior, as Woodward (2009) suggests. That is to say, what we need to establish is whether they share with behavioral experiments the same errors and biases with respect to motivations. Clearly not all neuroeconomics experiments will satisfy the required independence: if the affect driving the behavior in the experiment is not really caused by perceived unfairness, but by some other feature of the experimental setup, then the imaging data (fMRI pictures of affect-related activation) would likewise not be caused by social preferences but instead would be an artifact of an experimental bias shared by the neuroeconomics and the behavioral experiments. In this case, the social preference hypothesis would not screen off the behavioral and neuroimaging evidence, because the results would be effects of the common error.

This does not mean, however, that neuroeconomics experiments can never be independent in the right way. Independence can be achieved in several ways, such as by varying the (behavioral) experimental protocol or by increasing (causally) independent sources of variation. Concerning the latter, changing the subjects' response to the intended social cue (such as perceived unfair play) by directly modulating the affective responses—for example, by administering benzodiazepine (Gospic et al. 2011) or oxytocin (Baumgartner et al. 2008) to experimental subjects—reduces the chance that the observed behavior is caused by something other than the intended cue (since it is hypothesized to be mediated by affect-related motivational mechanisms). Adding further methods of data production, such as self-reporting, can provide further suitably independent, triangulating grounds: it would be unlikely, although certainly not impossible, for the subjects to report that they acted out of anger toward apparently unfair play by the other players if in reality their behavior was caused by some other feature of the experiment (such as pure anxiety or frustration arising from the laboratory environment) or from a mistaken, but fundamentally self-interested, learning strategy.

Nevertheless, such independence will not always be sufficient grounds for an inference from diverse evidence to the truth of a general theory (e.g., strong reciprocity theory vs. rational choice theory). It may still be the case that the convergence between neuroeconomics experiments and behavioral experiments increases confidence in, for example, the existence of social preferences as a robust laboratory phenomenon but that the kind of situations captured by one-shot games are never found in the wild (Woodward 2009) or that social preferences do not in fact sustain pro-social behavior where the general theory predicts they would (Guala 2012). More generally, even though diversity of evidence is sometimes cited as a powerful argument in favor of a general theory based on social preferences, such as strong reciprocity theory (e.g., Camerer 2013; Fehr and Krajbich 2014), the argument that a theory's ability to account for several findings speaks in favor of its truth should be kept distinct from the robust determination of a phenomenon from multiple sources of data. Our account of triangulation does precisely this.

6. Conclusion. We have argued that the intuitive idea of the epistemic gain of variety of evidence works differently at the level of inferences from phenomena to theory and from data to phenomena. We offered an account of triangulation understood as a matter of controlling for error in causal inference to make claims about phenomena and addressed some of the skeptical critiques leveled against triangulation. Applied to the causal data-to-phenomenon inferences, the confirmational independence condition captures the epistemic value of triangulation, and this condition is realized when the different methods are reliability independent. Can we infer from convergent results obtained by highly unreliable procedures the correctness of the result? Of course not. Nor does employing a different procedure suffice to ensure that every possible error has been eliminated. Nevertheless, these concerns should warn us against conferring epistemic superpowers to triangulation, not lead us to conclude that it is without merits. The philosophical task is to identify the conditions under which a concordance of results supports rational increases in confidence in the result, as we have illustrated in the case of bringing neuroeconomics evidence to triangulate on the existence and causal role of social preferences.

REFERENCES

Andreoni, James, and B. Douglas Bernheim. 2009. "Social Image and the 50–50 Norm: A Theoretical and Experimental Analysis of Audience Effects." *Econometrica* 77:1607–36.

Baumgartner, Thomas, Markus Heinrichs, Aline Vonlanthen, Urs Fischbacher, and Ernst Fehr. 2008. "Oxytocin Shapes the Neural Circuitry of Trust and Trust Adaptation in Humans." *Neuron* 59:639–50.

Binmore, Ken. 2005. "Economic Man—or Straw Man?" *Behavioral and Brain Sciences* 28: pt23–24.

Blaikie, Norman. 1991. "A Critique of the Use of Triangulation in Social Research." *Quality and Quantity* 25:115–36.
Bogen, James, and James Woodward. 1988. "Saving the Phenomena." *Philosophical Review* 47: 303–52.
Bovens, Luc, and Stephan Hartmann. 2002. "Bayesian Networks and the Problem of Unreliable Instruments." *Philosophy of Science* 69:29–72.
Camerer, Camerer. 2013. "Experimental, Cultural, and Neural Evidence of Deliberate Prosociality." *Trends in Cognitive Science* 17:106–7.
Campbell, Donald T., and Donald W. Fiske. 1959. "Convergent and Discriminant Validation by Multitrait-Multimethod Matrix." *Psychological Bulletin* 56:81–105.
Cartwright, Nancy. 1991. "Replicability, Reproducibility, and Robustness: Comments on Collins." *History of Political Economy* 23:143–55.
Clarke, Cristopher. 2014. "Neuroeconomics and Confirmation Theory." *Philosophy of Science* 81:195–215.
Claveau, François. 2011. "Evidential Variety as a Source of Credibility for Causal Inference: Beyond Sharp Designs and Structural Models." *Journal of Economic Methodology* 18:233–53.
Craver, Carl, and Anna Alexandrova. 2008. "No Revolution Necessary: Neural Mechanisms for Economics." *Economics and Philosophy* 24:381–406.
Fehr, Ernst, and Ian Krajbich. 2014. "Social Preferences and the Brain." In *Neuroeconomics: Decision Making and the Brain*, 2nd ed., ed. Paul W. Glimcher and Ernst Fehr, 193–218. London: Elsevier.
Fitelson, Branden. 2001. "A Bayesian Account of Independent Evidence with Applications." *Philosophy of Science* 68:123–40.
Franklin, Allan. 1986. *The Neglect of Experiment*. Cambridge: Cambridge University Press.
Fumagalli, Roberto. 2013. "The Futile Search for True Utility." *Economics and Philosophy* 29: 325–47.
Gospic, Katarina, Erik Mohlin, Peter Fransson, Predrag Petrovic, Magnus Johannesson, and Martin Ingvar. 2011. "Limbic Justice: Amygdala Involvement in Immediate Rejection in the Ultimatum Game." *PLOS Biology* 9:1–8.
Guala, Francesco. 2012. "Reciprocity: Weak or Strong? What Punishment Experiments Do (and Do Not) Demonstrate?" *Behavioral and Brain Sciences* 35:1–59.
Gul, Faruk, and Wolfgang Pesendorfer. 2008. "The Case for Mindless Economics." In *The Foundations of Positive and Normative Economics: A Hand Book*, ed. Andrew Caplin and Andrew Schotter, 3–39. Oxford: Oxford University Press.
Hacking, Ian. 1983. *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge: Cambridge University Press.
Hammersley, Martyn. 2008. "Troubles with Triangulation." In *Advances in Mixed Method Research*, ed. Manfred Bergman, 22–36. London: Sage.
Hey, Spencer. 2015. "Robust and Discordant Evidence: Methodological Lessons from Clinical Research." *Philosophy of Science* 82:55–75.
Hudson, Robert. 1999. "Mesosomes: A Study in the Nature of Experimental Reasoning." *Philosophy of Science* 66:289–309.
———. 2009. "The Methodological Strategy of Robustness in the Context of Experimental WIMP Research." *Foundations of Physics* 39:174–93.
———. 2013. *Seeing Things: The Philosophy of Reliable Observation*. Oxford: Oxford University Press.
Kuorikoski, Jaakko, Aki Lehtinen, and Caterina Marchionni. 2010. "Economic Modelling as Robustness Analysis." *British Journal for the Philosophy of Science* 61:541–67.
———. 2012. "Robustness Analysis Disclaimer: Please Read the Manual before Use!" *Biology and Philosophy* 27:891–902.
Kuorikoski, Jaakko, and Petri Ylikoski. 2010. "Explanatory Relevance across Disciplinary Boundaries: The Case of Neuroeconomics." *Journal of Economic Methodology* 17:219–28.
Levitt, Steven, and John List. 2007. "What Do Laboratory Experiments Measuring Social Preferences Reveal about the Real World?" *Journal of Economic Perspectives* 21:151–74.
Odenbaugh, Jay, and Anna Alexandrova. 2011. "Buyer Beware: Robustness Analyses in Economics and Biology." *Biology and Philosophy* 26:757–71.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.
Ross, Don. 2008. "Two Styles of Neuroeconomics." *Economics and Philosophy* 24:473–83.
Salmon, Wesley. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.
Schickore, Jutta, and Klodian Coko. 2013. "Using Multiple Means of Determination." *International Studies in the Philosophy of Science* 27:195–313.
Schupbach, Jonah. 2015. "Robustness, Diversity of Evidence, and Probabilistic Independence." In *Recent Developments in the Philosophy of Science: EPSA13 Helsinki*, ed. Uskali Mäki, Ioannis Votsis, Stéphanie Ruphy, and Gerhard Schurz, 305–16. New York: Springer.
Sober, Elliott. 1989. "Independent Evidence about a Common Cause." *Philosophy of Science* 56: 275–87.
Soler, Léna, Emiliano Trizio, Thomas Nickles, and William Wimsatt. 2012. *Characterizing the Robustness of Science*. Boston Studies in the Philosophy of Science. Dordrecht: Springer.
Staley, Kent W. 2004. "Robust Evidence and Secure Evidence Claims." *Philosophy of Science* 71:467–88.
Stegenga, Jacob. 2009. "Robustness, Discordance, and Relevance." *Philosophy of Science* 76:650–61.
———. 2012. "Rerum Concordia Discors: Robustness and Discordant Multimodal Evidence." In Soler et al. 2012, 207–26.
Weisberg, Michael. 2006. "Robustness Analysis." *Philosophy of Science* 73:730–42.
Wimsatt, William. 1981. "Robustness, Reliability, and Overdetermination." In *Scientific Inquiry in the Social Sciences*, ed. M. Brewer and B. Collins, 123–62. San Francisco: Jossey-Bass.
———. 2007. *Re-engineering Philosophy for Limited Beings*. Cambridge, MA: Harvard University Press.
Woodward, James. 2000. "Data, Phenomena, and Reliability." *Philosophy of Science* 67:163–79.
———. 2003. *Making Things Happen*. Oxford: Oxford University Press.
———. 2006. "Some Varieties of Robustness." *Journal of Economic Methodology* 13:219–40.
———. 2009. "Experimental Investigations of Social Preferences." In *The Oxford Handbook of Philosophy of Economics*, ed. Harold Kincaid and Don Ross, 189–222. Oxford: Oxford University Press.