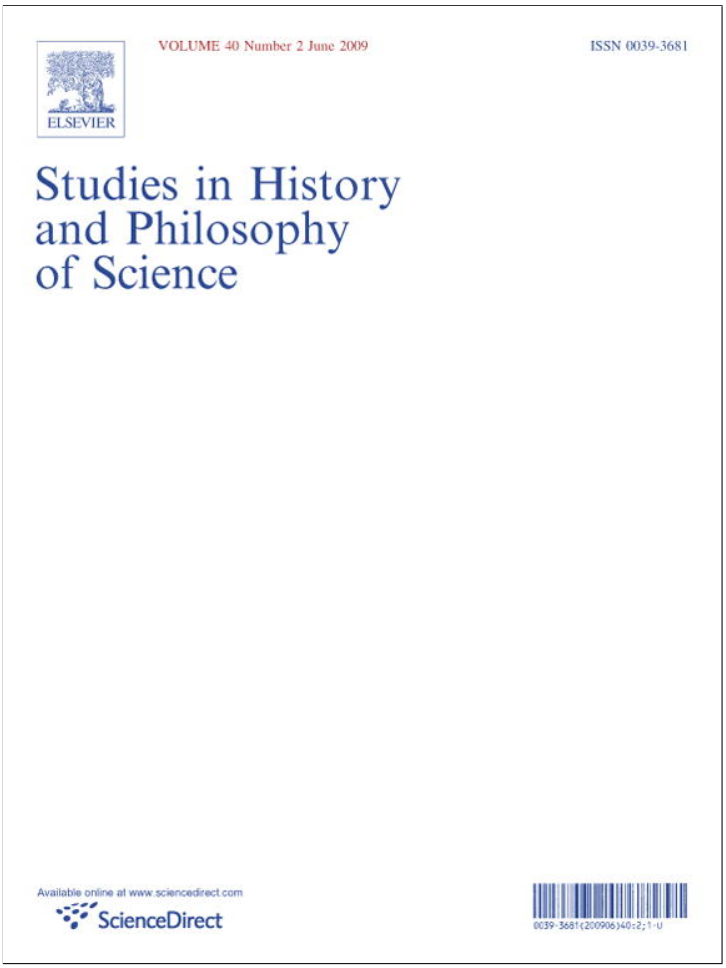


Studies in History and Philosophy of Science 40 (2009) 225–226

Contents lists available at ScienceDirect

Studies in History and Philosophy of Science

journal homepage: www.elsevier.com/locate/shpsa

Introduction

Philosophy in the trenches: from naturalized to experimental philosophy (of science)

Karola Stotz

Department of Philosophy, Quadrangle 14, University of Sydney, NSW 2006, Australia

ARTICLE INFO

Keywords:
Experimental philosophy
Naturalism
Continuity thesis
Philosophy of science
Quantitative data

ABSTRACT

Recent years have seen the development of an approach both to general philosophy and philosophy of science often referred to as 'experimental philosophy' or just 'X-Phi'. Philosophers often make or presuppose empirical claims about how people would react to hypothetical cases, but their evidence for claims about what 'we' would say is usually very limited indeed. Philosophers of science have largely relied on their more or less intimate knowledge of their field of study to draw hypothetical conclusions about the state of scientific concepts and the nature of conceptual change in science. What they are lacking is some more objective quantitative data supporting their hypotheses. A growing number of philosophers (of science), along with a few psychologists and anthropologists, have tried to remedy this situation by designing experiments aimed at systematically exploring people's reactions to philosophically important thought experiments or scientists' use of their scientific concepts. Many of the results have been surprising and some of the conclusions drawn from them have been more than a bit provocative. This symposium attempts to provide a window into this new field of philosophical inquiry and to show how experimental philosophy provides crucial tools for the philosopher and encourages two-way interactions between scientists and philosophers.

It is open to us to regard philosophy and science not as radically distinct modes of inquiry but rather as aspects of a single inquiry focused in somewhat differing ways, distinguished by their intellectual aims, and not truly by any restrictions on their 'proper' methods. A central locus of this continuity thesis within philosophy is in philosophy of science, where the drive to know a science intimately—in order to theorize it responsibly—plays a major role. Think of the way in which current philosophers of physics study QM or current philosophers of biology interact with important thinkers in the fields of population biology, ecology, development and molecular biology, to name just a few. Clearly this approach leads its practitioners to appreciate just how closely related philosophy and the science it applies to are. This constitutes a 'natural history' style approach to continuity: individual theorists pay exquisite care to a particular part of the scientific landscape in order to explicate it faithfully. Note, however, that this short introduction does not aim to sum up all other possible attitudes to the relation between science and the philosophy of science. Both the papers by Weinberg and Crowley and by Stotz discuss this topic in more depth, and see especially a recent special issue of the *Quarterly Review of Biology*, for example (Pigliucci, 2008, in particular the paper by Stotz & Griffiths, 2008).

Having embraced continuity, why has this segment of the philosophical community stopped at the level of the natural historical approach? Why restrict our interaction with science to exchanges based on only qualitative information—why not investigate the possibility of acquiring quantitative information as well? One might speculate about two reasons. First, we were afraid that if we adopted a more seriously continuous approach we'd lose our identity and become sociologists or historians—what would the difference be between a sociologist of science and a philosopher of science doing ethnographies? Worse still, we'd only qualify as exceedingly poor sociologists at that. This is no idle worry, for philosophers are inclined to conduct 'research', when they do, in an

E-mail address: karola.stotz@gmail.com

0039-3681/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.shpsa.2009.03.011

extremely sloppy fashion. Frank Jackson's (1998) observation that his intuitions about Gettier cases are confirmed by asking his students about their intuitions is a case in point (see Stich & Weinberg, 2001, for a 'continuist' critique along those lines). But it also points the way forward—since philosophers have more seriously begun doing this sort of thing why not do it well?

Recent years have seen the development of an approach both to general philosophy and philosophy of science which a number of people have been calling 'experimental philosophy', or 'X-Phi' for short. Philosophers often follow their intuition to make or presuppose empirical claims about how people would react to hypothetical cases and thought experiments. But their evidence for claims about what 'we' would say is usually very limited indeed. A growing number of philosophers, along with a few psychologists and anthropologists, have tried to remedy this situation by designing experiments aimed at systematically exploring people's reactions to philosophically important thought experiments. In a similar vain philosophers of science have gathered empirical data on how key scientific concepts are understood by particular scientific communities, or have compared the vernacular understanding of certain popular-scientific concepts between laymen and scientists (e.g. Griffiths & Stotz, 2008). Many of the results have been surprising and some of the conclusions drawn from them have been more than a bit provocative.

Topics of recent work in this field are free will, compatibilism and responsibility, the identity of objects, the nature of conceptual change in science, the vernacular understanding of the concept innateness, the role of scientific concepts in the reception and dissemination of scientific results, the application of scientific concepts to social situations (e.g. the naturalistic fallacy), the reference of proper names, moral realism and fundamental moral disagreement, epistemic norms and the concept of knowledge, and the role of appeals to intuition in philosophy.

Experimental philosophy then is the natural response to both of these earlier concerns. Its practitioners do not appear to have lost their identity as philosophers nor are they displaying fatal disregard for the methodological niceties of well conducted empirical research (though there is still a learning curve here to be navigated). The early flourishing of this way of doing philosophy suggests that we philosophers will benefit if we are willing to pursue continuity more seriously; it is in just those places where the line between philosophy and science is most blurred (e.g. cognitive science or the philosophy of biology) that much of the most exciting work in philosophy is occurring.

So how, then, does X-Phi exemplify this continuity between philosophy and science? First, X-Phi provides philosophers with *scientific* tools relevant to their own recognizably *philosophical* inquiries. Some of this work has been mostly methodological, issuing an empirically-based critique of philosophy in the 'separate but equal' tradition (Machery et al., 2004; Weinberg et al., 2001). But other work has been philosophically positive as well, ranging from explicating the role of affect in our moral psychology (Nichols & Knobe, 2007) or the moral/conventional distinction (Kelly et al., 2007) to attempts to adjudicate burdens-of-proof issues in free will (Nadelhoffer et al., 2005). Second, X-Phi exemplifies the new philosophy–science continuity by expanding the set of tools available to those philosophers of the sciences who seek a more active, engaged discourse with their target disciplines. Stotz and Griffiths's research, in collaboration with a large international group of philosophers and scientists, actively surveying the uses of 'gene', and more recently 'innateness', in different subdisciplines of the biosciences suggests an unprecedented level of quantitative, experimental interaction between biologists and the philosophers who study them. Moreover, the interaction runs in both directions: they do not merely analyze the biologists' 'gene' concepts, but make proposals as to how researchers can avoid confusion and fruitless misunderstandings by recognizing the variety of such concepts on offer. We would suggest that good candidates for a similar experimental treatment would include 'mental representation' in cognitive science, and indeed 'intuition' in philosophy.

While none of the participants argue that X-Phi is *the* method for approaching traditional philosophical questions, or even *the* method for a naturalistically-committed philosophy, it is, importantly, *a* method for both, and its successes illuminate the interpenetrating relationship between philosophy and the sciences. The three papers present a representative window into this new field: Weinberg and Crowley analyze the status of X-Phi as a mode of philosophical inquiry and critically reflect on the virtues and vices of experimental philosophy and epistemology. Stotz presents a hybrid paper (partly research report, partly meta-reflection) about the 'Representing Genes' study that discusses the broader epistemological framework within which that research was conducted, and reflects on the relationship between science, history and philosophy of science, and society. The last paper presents a research report: Knobe presents data from some recent experiments showing how people's causal judgments are affected by moral considerations and offers an explanation for this phenomenon.

References

Griffiths, P. E., & Stotz, K. (2008). Experimental philosophy of science. *Philosophy Compass, 3*(3), 507–521.

Jackson, F. C. (1998). *From metaphysics to ethics: A defence of conceptual analysis*. Oxford: Oxford University Press.

Kelly, D., Stich, S., Haley, K., Eng, S., & Fessler, D. (2007). Harm, affect and the moral/conventional distinction. *Mind & Language, 22*(2), 117–131.

Machery, E., Mallon, R., Nichols, S., & Stich, S. P. (2004). Semantics, cross-cultural style. *Cognition, 92*(3), B1–B12.

Nadelhoffer, T., Nahmias, E., Morris, S., & Turner, J. (2005). Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology, 18*(5), 561–584.

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs, 41*, 663–685.

Pigliucci, M. (Ed.). (2008). Science and philosophy symposium. *Quarterly Review of Biology, 83*(1), 3–86. (Special issue)

Stich, S., & Weinberg, J. (2001). Jackson's empirical assumptions. *Philosophy and Phenomenological Research, 62*(3), 637–643.

Stotz, K., & Griffiths, P. E. (2008). Biohumanities: Rethinking the relationship between biosciences, philosophy and history of science, and society. *Quarterly Review of Biology, 83*(1), 37–45.

Weinberg, J., Nichols, S., & Stich, S. (2001). Normativity and epistemic intuitions. *Philosophical Topics, 29*, 429–460.

Studies in History and Philosophy of Science 40 (2009) 227–232

Contents lists available at ScienceDirect

Studies in History and Philosophy of Science

journal homepage: www.elsevier.com/locate/shpsa

The x-phi(les): unusual insights into the nature of inquiry

Jonathan M. Weinberg [a], Stephen Crowley [b]

[a] Department of Philosophy, Indiana University, Sycamore Hall 026, Bloomington, IN 47405, USA
[b] Department of Philosophy, Boise State University, 1910 University Dr., Boise, ID 83725, USA

ARTICLE INFO

Keywords:
Science
Philosophy
Experimental Philosophy

ABSTRACT

Experimental philosophy is often regarded as a category mistake. Even those who reject that view typically see it as irrelevant to standard philosophical projects. We argue that neither of these claims can be sustained and illustrate our view with a sketch of the rich interconnections with philosophy of science.

1. Introduction

Experimental philosophy (henceforth 'XΦ') takes seriously the idea that philosophical inquiry may benefit directly from quantitative empirical research. That view strikes many as deeply misguided, perhaps oxymoronic: experimentation is simply the wrong *kind* of investigatory technique for solving philosophical puzzles. But to think XΦ an oxymoron is to have an opinion about the relationship between scientific and philosophical inquiry—minimally, that philosophy and science are distinct, independent enterprises each pursuable on its own terms. We will suggest here, by way of putting the question on the table, that neither this 'separate but equal' view nor any other global account of *the* relation between science and philosophy can be maintained. This however leaves open the possibility that XΦ may be distinguished from philosophy 'proper' in a more local fashion. Having sketched our reasons for being skeptical about the possibility of a global relation between scientific and philosophical inquiry, we move on to give more careful consideration to the possibilities for local demarcation, arguing that XΦ itself cannot be separated off from philosophy more generally. We then consider how XΦ might 'play well with others', with a brief discussion of the possibilities for developing a relationship between XΦ and philosophy of science.

Following Kornblith (1994) we identify three possible relations between scientific and philosophical inquiry. First, the two approaches may be seen as direct competitors, that is, science and philosophy constitute alternative approaches to a single explanatory task. Since at most one such alternative can be correct with the other discarded as a failure we label this view the *replacement* thesis. A second approach, hinted at above, sees science and philosophy as pursuing distinct explanatory goals with distinct methodologies. This 'separate but equal' approach we will label the *traditional* approach. The third option is that science and philosophy constitute generally overlapping areas of inquiry beneath the umbrella of inquiry in general. This view we call the *continuity* hypothesis.[1]

On which of these views does XΦ turn out automatically to be an impossible project? Not on the continuity view. On the replacement view, for example Quine (1969), *all* of philosophy becomes impossible (appearances to the contrary), so that would do the trick; we suspect however that this is not the view that most critics of XΦ have in mind. It is the *traditional* 'separate but equal' view that, we think, lies behind most philosophers' complaints in this regard. If it is plausible that 'separate but equal' is an accurate characterization of the relation of philosophy and science, then it would follow that it is also plausible that XΦ can be preemptively dismissed as an illicit combination of the two projects.

E-mail addresses: jmweinbe@indiana.edu (J.M. Weinberg); stephencrowley@boisestate.edu (S. Crowley)

[1] Precisely what *continuity* amounts to is a complex issue. It certainly covers the kind of work (e.g. Kelly et al., 2007) that is of equal interest to philosophers and scientists. It may or may not also cover the kind of science–philosophy relation described in Friedman's (2001) *Dynamics of reason* where the two fields interact only under very unusual circumstances. Fortunately for us, the precise nature of *continuity* does not affect the arguments we develop in what follows.

0039-3681/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.shpsa.2009.03.016

But philosophy's conceptual geography—both currently and historically—simply does not underwrite a general adherence to 'separate but equal'. Indeed, we find it doubtful that *any* of these models can apply to the philosophy–science relationship on the whole. For in some places, replacement seems to be the historical fact, and unproblematically so: we hope it needs no special argument on our part that it is a good thing that much of what had been natural philosophy now is just physics, or biology, or chemistry. In other places, it seems plausible that 'separate but equal' is unproblematic—logic, for example (though a story would need to be told about the particular nature of logic's relations with, say, computer science). Finally there are yet other parts of philosophy for which the continualist approach seems both descriptively accurate and normatively appropriate; for example, large parts of the contemporary philosophy of mind run together with the theoretical psychology literature. Just to take a few obvious examples from the last two decades, one might consider Fodor and Pylyshyn (1988)—a philosopher and a psychologist—on cognitive architecture and connectionism, Griffiths (1997) on emotions, Cowie (1999) on innateness, Block (2001) on the neural correlates of consciousness, Allen (2006) on animal cognition, or Nichols and Stich (2003) on folk psychology. The same also seems true of some parts of philosophy of language and linguistics, such as Jackson (1987) on conditionals, Recanati (2000) on indirect reference, Devitt (2006) on linguistic knowledge, or, for that matter, any of a number of works by Chomsky himself, such as the papers in Chomsky (2000).

In sum, it does not seem to us that there can be any quick, general answer to the question, is XΦ a genuinely possible form of inquiry? The relations between philosophy and science are multivarious and local; so too must our question about the status of XΦ be one that can only be given a local answer. In the next section, we will argue that no appropriately local line can be drawn that will put all of XΦ on one side of a science/philosophy distinction, without taking large chunks of non-XΦ philosophy with it.

2. Can we substantively demarcate XΦ from 'traditional' philosophy?

The development of XΦ could be seen as challenging more traditional forms of intuition-based philosophy, much as the development of careful quantitative methods challenged purely observational 'naturalistic' methods in biology centuries ago. One way traditional philosophers might seek to preserve their investigatory methods would be to show that one can cleanly partition off XΦ from standard analytic philosophy, as simply doing something different that need not be viewed as a competitor any more than formal mathematics need view empirical science as a competitor. Indeed, we take this to be the modest, reasonable position at the heart of the sometimes hyperbolic talk of XΦ's being incoherent on its face. Although we do not seek to push this challenge here,[2] we are concerned that this defensive maneuver is based on a mistaken view of the relations between XΦ and philosophy more generally. We consider three different strategies by means of which a practitioner of standard analytic philosophy (henceforth 'SAP') might seek to draw a bright line between SAP and XΦ, and argue that none is successful.

First, and perhaps most obviously, it might be suggested that the two are, after all, robustly distinct methods, or at least families of methods. They wear this difference on their nomenclatural sleeves: the one is *analytic*, and the other is *experimental*. Once we identify the core set of methodological norms that give them their identities, then we can keep them cleanly separated.

However, on close inspection the various researchers working under the XΦ umbrella have less in common than might be suspected. One can locate differences between their methods along a number of fundamental dimensions, and with these different methods in view, it will be impossible to draw a line that puts all of them on one side, but all of SAP on the other. We will articulate a few suggestive examples here, indicating their variation along such dimensions as subject population, explanatory orientation, and investigatory goals.

At one extreme, we find Joshua Knobe attempting to explore the content of various folk-psychological concepts such as intentional action (Knobe, 2006) and valuing (Knobe & Roedder, Forthcoming). One of his chief investigatory interests has been to argue that our folk-psychological concepts are not solely structured for the purposes of prediction and explanation, but also for moral evaluation and the assignment of praise and blame. His target population, therefore, is 'the folk' at large, frequently college undergraduates because they are cheap and plentiful, but also on occasion people in the street, for example, Knobe (2003). His inferential strategy for getting at the content of our folk concepts is what we will call *ethological*: the relevant facts are more or less read off of the behavior itself, without an ambition to articulate the underlying psychological mechanisms that produce that behavior.[3] The relevant philosophical behavior is the expressed categorical decisions of the subjects to hypothetical cases—in many ways a paper-and-pencil version of the fundamental evidential tool of SAP. This basic model applies to several other practitioners of XΦ such as Thomas Nadlehoffer and Eddy Nahmias (see Nahmias et al., 2005, and Nadelhoffer & Nahmias, 2005, 2007), though their interest has generally been more in determining what the intuitions of 'the folk' are for the purposes of refereeing burden-tennis matches in philosophy than in the predictive or moral components of folk psychological concepts more generally.

Karola Stotz and Paul Griffiths (see Stotz & Griffiths, 2004, and Stotz et al., 2004) are also trying to map the content of concepts—the various gene concepts at play in different communities of biologists—and as such their methods share some similarities with Knobe's. They also present their subjects with surveys about hypothetical cases, albeit ones that have significant real-world plausibility, and they also have an ethological approach to the data so gathered, where the interest is more in the distribution of different patterns of categorizations than in the underlying psychological mechanisms. But their larger theoretical goals are different. Stotz and Griffiths are pursuing the concepts of different communities of experts, not the folk, and where the others work from a starting presupposition that the categories of interest to them are broadly, even universally shared, Stotz and Griffiths are particularly interested in documenting the *differences* across those communities. Indeed, they bring to bear not just the survey methods of social psychology, but the tools of population ecology as well, to help map out the 'intellectual ecology' of the different 'subspecies' of gene concepts.

Not all XΦ practitioners are ethological, however. Stephen Stich, Shaun Nichols, Ron Mallon (among others) have pursued the *psychological* questions of the processing architecture—the 'boxology'—of various of our cognitive capacities (see Kelly et al., 2007,

[2] Except for *very* briefly in our concluding section.

[3] Although a certain amount of psychological theorizing is required to help distinguish between some relevant hypotheses (for example, to fend off attempts to explain the data away as the artifact of the pragmatics of the survey items) the main focus remains descriptive.

or Nichols & Mallon, 2006). This work often enters into direct dialogue with research conducted by psychologists and published in psychological forums (e.g. Turiel, 1979, or Nucci, 2001).

A further XΦ project, one which does not share the basically explanatory orientation of the research just mentioned, has more pointedly polemical ambitions. This project aims to demonstrate that intuitions of the sorts typically appealed to in SAP diverge across various demographic lines, and indeed may not even be stable within individual persons (see Weinberg et al., 2001; Swain et al., 2008; Machery et al., 2004; Nichols & Ulatowski, 2007; for an overview, see Alexander & Weinberg, 2007). This work is mostly predicated on the plausibility of generalizing from the results with their subjects (again, mostly undergraduates) to the patterns of intuitive judgments of philosophers, although some researchers are carrying out direct studies of professional philosophers as well.

So there is nothing like a rich, deeply shared method that holds across the whole XΦ community. At best one might suggest that they all use surveys, but that will not really serve to distinguish it from SAP, when SAP relies on fundamentally the same kind of data: namely, the categorical judgments of their 'subjects'. And this appearance of a minimally shared method across all XΦ is perhaps more misleading than revealing; there is important work being done using brain imagining technologies as well, for example, Greene & Haidt (2002), and with historical/genealogical approaches, for example, Nichols (2007). Furthermore, different parts of XΦ share substantive goals with SAP, such as coming to understand the structure and nature of various philosophically interesting concepts, or evaluating the kind of justification that intuition may (or may not) provide. There are *some* differences between the methods of XΦ and SAP, of course, but they do not appear sufficient to establish a stark discontinuity between them.

Nonetheless, it might be contended that there is another important difference between SAP and XΦ. The latter has *empirical commitments* while the former is unsullied by the empirical world. For SAP is meant to be modally neutral, presupposing nothing about the particular possible world we are in; whereas XΦ, by its very nature, draws upon specific empirically-ascertained—and hence contingent—data. In addition to its prima facie plausibility, this move also has the benefit of not seeming entirely ad hoc: SAP has ambitions of providing deliverances with necessary modal force, for example, Jackson (1998), but XΦ's contingent premises doom it to merely contingent conclusions.[4] Given SAP's goals, then, its practitioners may have good reason to hold XΦ at arm's length, or longer.

Now, this tactic only works if the methods of SAP are indeed devoid of entanglements with contingency. But a lack of empirical *premises* is not the same as a lack of empirical *commitments*. We know that scientific instruments embody empirical information about the world, for example, Baird (2004). And this principle can be extended from material instruments to practices more generally. For example, the standard practice in Chomskyan linguistics of preferencing the intuitions of linguistically naïve native speakers is only sensible given a broad set of presuppositions about the general modularity of syntactic knowledge; the thinness of how that knowledge can be made manifest to consciousness (and hence reportable); and the possibility that explicit linguistic belief can interfere with such knowledge.

Furthermore, although practitioners are often aware of some of their most important presuppositions, this is not always so. Timberlake (2002) demonstrates that the learning paradigms of traditional behaviorists, without their realizing it, embody significant information about the niches that their animal subjects function well in, and very particular facts about how they so function. The training apparatus for any given type of animal will be 'tuned' to that animal, so that the researcher can produce enough appropriate behavior from the subjects to do their intended study. As a result of this process of bit-by-bit conforming the apparatus to the animal:

> it is possible to infer from their standard Skinner boxes that pigeons forage for grain based primarily on pecking related to visual stimuli, do not like to move about in the dark, and are not inclined to feed out of holes they cannot see into. Rats ... use manipulation responses, attend to movement and sound, are less fond of brightly lighted areas, and have no compunction about poking into dark holes for food. (Ibid., p. 357)

The behaviorists, of course, do not take themselves to be including so much species- and niche-specific information in their studies—they are after general laws of learning, and it is inconsistent with their theoretical goals to explain animal behaviors in terms of individual species-natures and suitability to distinct environments and behaviors. This information became embodied, rather, over the course of trying to get a working apparatus. Particular empirical information therefore can be built in to a set of practices or instruments not only without the practitioners' awareness of it, but even against their express desires.

We contend that SAP, too, has its share of empirical commitments. To escape Meno's paradox, analytic philosophers must treat the sources of intuition as being cognitively distinct from our explicitly held theories. Also, one's explicit theory can sometimes interfere with one's intuitions, which is why occasionally a rival theorist's intuitions may be explained away as being 'contaminated' by their preferred theory. (It would be a nice piece of history to determine to what extent this move in philosophical practice antedated or postdated the rise of concerns about theory-ladenness of observation.)

Moreover, SAP is committed to the well-functioning of intuition across arbitrary stipulation. If you have two rival analyses—say, *that all As are F* and *that all As are G*—then the strategy is to devise a hypothetical situation, in which there is a relevant presence of F and absence of G (or vice versa), and check our intuitions for A-ness. If F and G tend overwhelmingly to travel together in our ordinary lives, then the target scenario may need to be rather bizarre or esoteric. And frequently, they are. (See, for example, Lehrer's gypsy lawyer, or Lehrer and Cohen's new evil demon, or Lackey's double-lesioned truth-telling liar: Lehrer, 1971; Lehrer & Cohen, 1983; Lackey, 2006.)

A third contingent component of SAP is that the intuitions cited as evidence are widely shared, intra- and inter-personally. One linguistic sign of this commitment is the common deployment of either a definite article ('the intuition') or the first-person plural possessive ('our intuition'). Moreover, SAP practitioners frequently *need* to have this commitment, for it is unclear what evidential force such intuitions could have, if they varied with philosophically-irrelevant factors like the ethnicity of the intuiter, or the order in which cases are considered. Sometimes philosophers seem to require the agreement of the folk at large; other philosophers only seem to demand the agreement of other well trained philosophers, but then they take on the additional commitment that

[4] We note that XΦ has not shown any interest in Kripke's necessary a posteriori.

philosophical training produces persons whose intuitions partake of a greater degree of truth-tracking oomph.[5]

There may be further empirical commitments than these, but these three are enough to secure our basic point. SAP is not empirically pure, and cannot be shielded from XΦ by an appeal to the latter's mucking about in the scientific world of observable contingencies.

Now, even while granting that SAP has empirical *presuppositions*, one may still object that SAP has no empirical *premises*; and the same cannot be said about XΦ. This objection's main claim is not obviously true—claims about 'the' or 'our' intuition about a certain case may be best understood as an empirical claim about how subjects at large, or at least philosophers, will treat that case. Yet even if it is true, it will not suffice to insulate SAP from an XΦ invasion, for empirical presuppositions may still be scientifically evaluated and, if the science goes the wrong way, invalidated. Moreover, being a responsible investigator surely includes appropriate openness to evidence concerning the state of one's apparatuses and practices, and whether their commitments are satisfied by the world, or not. So, although the premise/presupposition distinction might be a good *epistemological* one, it is not clearly a useful *methodological* one.

So SAP and XΦ cannot be seen as discontinuous either in terms of their methods or of their having empirical commitments. One last possible source of discontinuity that we will briefly consider is historical. Perhaps they belong to different 'intellectual clades'—the philosophical on the one hand, the scientific on the other, for example—and thus are informed by and responsible to different traditions, having inherited different sets of problems that they seek to engage in and advance. It would thus be reasonable for SAP and XΦ to operate in parallel, non-competing, separate spheres.

Yet such an argument would get the particular history of XΦ wrong, as well as the broader history of the relation between philosophy and science. The aforementioned different research groups in XΦ all come from varying traditions of inquiry, all of them recognizably philosophical. The first, empirical conceptual analysis group, descends from analytic metaphysics itself; the second such group, from philosophy of science (see below); the third, psychological group from the kind of empirically-informed philosophy of cognitive science that has been active at least since the origins of cognitive science about a half-century ago; the fourth, critical group from the empiricist, anti-metaphysical tradition in epistemology that stretches honorably back at least to Hume. The similarities that are observed between the different forms of XΦ are perhaps better seen as analogies, not homologies. Moreover, there is a long philosophical tradition of entwining one's philosophy with one's psychology, and as Knobe has observed, XΦ largely examines 'the sorts of questions one finds in the work of Plato, Aristotle, Spinoza, Hume, Nietzsche, and so many others' (Knobe, 2007, p. 121). It will thus be impossible to find an ancestor on philosophy's evolutionary tree such that all of XΦ is found along one descending branch, without that branch including much of SAP itself.

3. Intersections: XΦ and philosophy of science

Having made the case for XΦ as a member in good standing of the philosophical community we can now consider how it might 'play with others' to the benefit of both parties. We illustrate this possibility with a brief discussion of how we see the relationship between XΦ and philosophy of science developing.[6]

The broad outlines of philosophy of science's contribution to XΦ are clear. To pursue XΦ is to deploy scientific methods in the service of philosophy. As a result, philosophy of science constitutes a source of expertise that experimental philosophers may draw upon in the pursuit of their inquiries.[7] For example, important progress in XΦ is sometimes made when a team of researchers notices that a key norm of experimental design has not been scrupulously observed by rival philosophers or psychologists. One can see Nichols and Ulatowski (2007) on the importance of within-subject designs and the danger of over-reliance on population means, or Kelly et al. (2007) on not letting one's instruments dictate the form that one's theories should take.

Since XΦ has relied heavily on survey instruments, philosophy of science can be of service in helping XΦ discern the appropriate domain for using such tools. While much information of interest to philosophers may be accessed using survey techniques that need not always be the case. For example, if you think emotions are important parts of moral judgments (e.g. Nichols, 2004), then the use of surveys may well be 'ecologically' inappropriate—that is, fail to generate the circumstances that allow for an accurate report of moral judgments.

Even more important than philosophy of science's contributions to XΦ, though, is XΦ's potential to expand the tools available to philosophy of science itself. The work of Stotz and Griffiths on the variety of gene concepts has been mentioned already (see above, Stotz & Griffiths, 2004, and also Stotz et al., 2004). Their methodology has much to offer other philosophers of science. Current philosophy of science cannot be faulted for the careful attention its practitioners pay to the details of the sciences they study. The insights of individual practitioners however are rarely quantified to even a superficial degree. These circumstances might with justice be compared to the glory days of Natural History where rural curates the length and breadth of Great Britain published richly detailed, insightful accounts of their local flora and fauna. Nonetheless this outpouring of detail led to little intellectual progress until the practices of quantification began to provide a framework within which the individual accounts could be compared and contrasted. The quantificational methodology of Stotz and Griffiths, we suggest, holds out the hope of similar progress within the philosophy of science. Surely this is a prospect desirable, exciting and worth working for.

A quite different direction is which XΦ may be taking philosophy of science is illustrated by the work of Stich and Kelly (Kelly et al., 2007). The point we wish to focus on here is the degree to which Stich and Kelly's work can be seen as not merely observing the psychologists' debate about moral/conventional distinction,

[5] Some neo-rationalists may avoid this particular flavor of empirical commitment, by endorsing a view in which only some intuition-like cognitions, those that meet a very epistemically rich set of conditions, count as real intuitions. (See, for example, BonJour, 1998; Bealer, 2000.) It may be that the vast majority of the folk—or even of philosophers—fail to have cognitions that meet such conditions, and so no such broad agreement will necessarily be expected. (Our thanks to an anonymous referee from this journal for pointing out this possibility.) We would contend, however, that such a position does not help SAP evade the general taint of contingency, for now practitioners of SAP face the contingent question: is *this* cognition that I am appealing to an epistemically-potent real intuition, or just an epistemically-dubious ersatz? (See Feltz, 2008, for further development of such arguments.)

It is important, in evaluating the relationship between SAP and XΦ, that one keep in mind that having a successful *epistemology* of intuitions does not entail that one also has a successful *methodology* of the deployment of such intuitions in wide philosophical practice. Nothing in our arguments here should be taken as in any way speaking to the first-order epistemological questions concerning the a priori status of individual intuitions, but only to these separate methodological questions.

[6] A topic of increasing interest—see When scientists and philosophers talk to each other: Proceedings of the Stony Brook 'SCI-PHI' Symposium (2008) (the symposium was organized by Massimo Pigliucci).

[7] We note in passing that the deployment of these methods by XΦ provides additional data on features of these methods for philosophers of science. Space precludes further discussion of this point here.

but actually participating in that debate. This approach stands in contrast to an important trend in much recent philosophy of science which supports, by default, a version of the 'separate but equal' doctrine regarding scientific and philosophical inquiry. It does so because it cedes ultimate intellectual authority to the science being studied. That is, although the science can (and does) impact the philosophy, the philosophy does not likewise impact the science.[8] That Stich and Kelly's work is violating this stricture (as is not uncommon in the philosophy of psychology) is, we submit, both a further example of a locus of continuity between philosophy and science, *and* indicative of a future for philosophy of science as a participant rather than a mere observer in scientific debates. Such a future need not be for everyone, but its existence surely provides a new and genuinely exciting opportunity for philosophers of science.

The possibilities highlighted by the work of Stotz, Griffiths, Stich and Kelly should not be taken to exhaust XΦ's capacity to contribute to philosophy of science. XΦ is a research project in its infancy and we are convinced that its contributions to philosophy of science (and philosophy more generally) will be limited only by the imaginations of its practitioners for the foreseeable future.

4. Briefly polemical conclusion

The moral of our story is that XΦ should not be seen as some alien invader, but as an organic outgrowth of philosophy itself, with significant promise for philosophy at large and also for the philosophy of science. One might well wonder, then, if there is *anything* that distinguishes XΦ from the kinds of traditional philosophy practised by proponents of SAP. Although we think there is no deep philosophical difference, especially as both do have (we argued) significant empirical commitments, they nonetheless have *different* empirical commitments. One consequence of this fact is that it is possible for the empirical commitments of one to be falsified, even while those of the other remain intact. And the aim of at least one camp of experimental philosophers has been to show, indeed, that there is a growing body of worrisome evidence for the falsity of SAP's presuppositions. Just because we have all descended from the same intellectual ancestors, it doesn't mean that we're all viewing the world from equally high points of the contemporary evidentiary fitness landscape. But that is perhaps an argument for another time.[9]

Acknowledgements

We would particularly like to thank Karola Stotz and the other participants in the symposium Philosophy in the Trenches: From Naturalized to Experimental Philosophy (of Science) at the 2006 Philosophy of Science Association Meeting in Vancouver, BC, and also the members of the Indiana University Experimental Epistemology Laboratory.

References

Alexander, J., & Weinberg, J. (2007). Analytic epistemology and experimental philosophy. *Philosophy Compass*, *2*(1), 56–80. (Available at doi:10.1111/j.1747-9991.2006.00048.x)

Allen, C. (2006). Transitive inference in animals: Reasoning or conditioned associations? In S. Hurley, & M. Nudds (Eds.), *Rational animals?* (pp. 175–185). Oxford: Oxford University Press.

Baird, D. (2004). *Thing knowledge: A philosophy of scientific instruments*. Berkeley: University of California Press.

Bealer, G. (2000). A theory of the a priori. *Pacific Philosophical Quarterly, 81*, 1–30.

Block, N. (2001). Paradox and cross-purposes in recent work on consciousness. *Cognition, 79*, 179–219.

BonJour, L. (1998). *In defense of pure reason*. Cambridge: Cambridge University Press.

Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. Oxford: Oxford University Press.

Chomsky, N. (2000). *New horizons in the study of language and mind*. Cambridge: Cambridge University Press.

Cowie, F. (1999). *What's within? Nativism reconsidered*. Oxford: Oxford University Press.

Devitt, M. (2006). *Ignorance of language*. Oxford: Oxford University Press.

Feltz, A. (2008). Problems with the appeal to intuition in epistemology. *Philosophical Explorations, 11*, 131–141.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*, 3–71.

Friedman, M. (2001). *Dynamics of reason*. Stanford: CSLI.

Giere, R. (1988). *Explaining science: A cognitive approach*. Chicago: University of Chicago Press.

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences, 6*(12), 517–523.

Griffiths, P. (1997). *What emotions really are: The problem of psychological categories*. Chicago: University of Chicago Press.

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge: Cambridge University Press.

Jackson, F. (1987). *Conditionals*. Oxford: Oxford University Press.

Jackson, F. (1998). *From metaphysics to ethics: A defense of conceptual analysis*. Oxford: Oxford University Press.

Kelly, D., Stich, S., Haley, K., Eng, S., & Fessler, D. (2007). Harm, affect, and the moral/conventional distinction. *Mind & Language, 22*(2), 117–131.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis, 63*, 190–193.

Knobe, J. (2006). The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies, 130*, 203–231.

Knobe, J. (2007). Experimental philosophy and philosophical significance. *Philosophical Explorations, 10*, 119–122.

Knobe, J., & Roedder, E. (Forthcoming). The ordinary concept of valuing. *Philosophical Issues*.

Kornblith, H. (1994). What is naturalistic epistemology? In idem (Ed.), *Naturalizing epistemology* (2nd ed.) (pp. 1–14). Cambridge, MA: MIT Press.

Kuhn, T. (1996). *The structure of scientific revolutions* (3rd ed.). Chicago: University of Chicago Press.

Lackey, J. (2006). Learning from words. *Philosophy and Phenomenological Research, 73*, 77–102.

Lehrer, K. (1971). How reasons give us knowledge, or, the case of the gypsy lawyer. *Journal of Philosophy, 68*(10), 311–313.

Lehrer, K., & Cohen, S. (1983). Justification, truth, and coherence. *Synthese, 55*, 191–207.

Machery, E., Mallon, R., Nichols, S., & Stich, S. (2004). Semantics, cross-cultural style. *Cognition, 92*, B1–B12.

Nadelhoffer, T., & Nahmias, E. (2005). Surveying freedom: folk intuitions about free will and moral responsibility. *Philosophical Psychology, 18*(5), 561–584.

Nadelhoffer, T., & Nahmias, E. (2007). The past and future of experimental philosophy. *Philosophical Explorations, 10*, 123–149.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2005). Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology, 18*(5), 561–584.

Nichols, S. (2004). *Sentimental rules: On the natural foundations of moral judgment*. New York: Oxford University Press.

Nichols, S. (2007). The ride of compatibilism: A case study in the quantitative history of philosophy. *Midwest Studies in Philosophy, 31*, 260–270.

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

Nichols, S., & Stich, S. (2003). *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds*. Oxford: Oxford University Press.

Nichols, S., & Ulatowski, J. (2007). Intuitions and individual differences: The Knobe effect revisited. *Mind & Language, 22*, 346–365.

Nucci, L. (2001). *Education in the moral domain*. Cambridge: Cambridge University Press.

Quine, W. V. O. (1969). Epistemology naturalized. In idem, *Ontological relativity and other essays* (pp. 69–90). New York: Columbia University Press.

Recanati, F. (2000). *Oratio obliqua, oratio recta*. Cambridge, MA: MIT Press.

Stotz, K., & Griffiths, P. (2004). Genes: Philosophical analyses put to the test. *History and Philosophy of the Life Sciences, 26*, 5–28.

Stotz, K., Griffiths, P., & Knight, R. (2004). How biologists conceptualize genes: An empirical study. *Studies in History and Philosophy of the Biological and Biomedical Sciences, 35*, 647–673.

Swain, S., Alexander, J., & Weinberg, J. (2008). The instability of philosophical intuitions: Running hot and cold on Truetemp. *Philosophy and Phenomenological Research, 76*, 138–155.

Timberlake, W. (2002). Niche-related learning in laboratory paradigms: The case of maze behavior in Norway rats. *Behavioural Brain Research, 134*, 355–374.

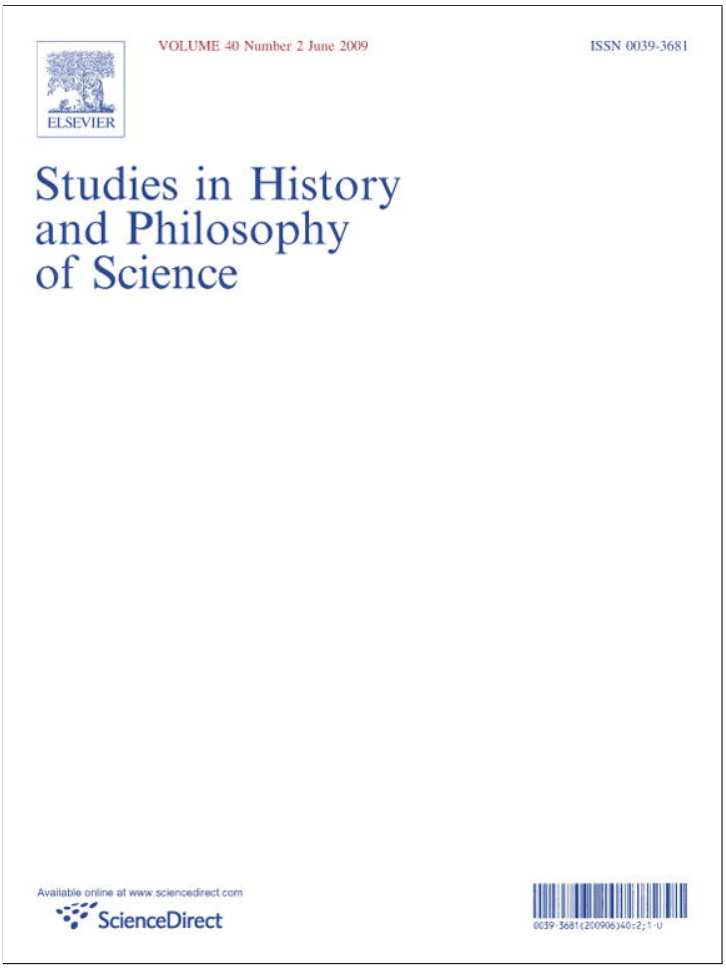
[8] For example, Ron Giere sees his task as being to 'explain the phenomenon of science itself in roughly the way that scientific theories explain other natural phenomena' (Giere, 1988, p. 1). Thus Giere gives science the same kind of control over philosophy of science that facts have over scientific theories. Similar sentiments are widespread in post-Kuhnian (Kuhn, 1996) philosophy of science (e.g. Hacking, 1983, or Chang, 2004, amongst many others that can be found on the author's bookshelves).

[9] For a review, see Alexander & Weinberg (2007).

Turiel, E. (1979). Distinct conceptual and developmental domains: Social convention and morality. In H. Howe, & C. Keasey (Eds.), *Nebraska Symposium on Motivation, 1977: Social cognitive development* (pp. 77–116). Lincoln, NE: University of Nebraska Press.

When scientists and philosophers talk to each other: Proceedings of the Stony Brook 'SCI-PHI' Symposium. (2008). *The Quarterly Review of Biology*, *83*(1), 3–86.

Weinberg, J., Nichols, S., & Stich, S. (2001). Normativity and epistemic intuitions. *Philosophical Topics, 29*, 429–460.



Studies in History and Philosophy of Science 40 (2009) 233–237

Contents lists available at ScienceDirect

Studies in History and Philosophy of Science

journal homepage: www.elsevier.com/locate/shpsa

Experimental philosophy of biology: notes from the field

Karola Stotz

Department of Philosophy, Quadrangle 14, University of Sydney, NSW 2006, Australia

ARTICLE INFO

Keywords:
Experimental philosophy
Biohumanities
Representing Genes Project
Gene concept
Science criticism
Conceptual ecology

ABSTRACT

I use a recent 'experimental philosophy' study of the concept of the gene conducted by myself and collaborators to discuss the broader epistemological framework within which that research was conducted, and to reflect on the relationship between science, history and philosophy of science, and society.

1. Introduction: experimental philosophy of biology and the 'biohumanities'

A central tenet underlying the work to be described here is that philosophy and science are not clearly separated activities, but aspects of a single inquiry into nature. They are distinguished primarily by the questions they ask, rather than by any restrictions on their 'proper' methods. The new field of 'experimental philosophy' (X-phi) pays tribute to this 'continuity thesis' by bringing empirical work to bear on philosophical questions. Its practitioners have not lost their identity as philosophers through their employment of methods traditionally associated with the sciences; rather experimental philosophers ('X-philes') perform experiments in an attempt to discern facts of relevance to *philosophical debates*. It is part of the burden of such work to show that some philosophical issues turn on an empirical supposition that may in turn be tested (Griffiths & Stotz, 2008). With respect to studying the changing concept of the gene previous research by the author and her collaborators has established that it is possible to operationalize questions about conceptual variation in a survey instrument and that the statistical analysis of these questionnaire data reveals the prevalence of particular gene concepts in different biological fields (Stotz et al., 2004). The Representing Genes Project at the center of this paper represents an extension of that earlier work.

The paper reflects on the motivation for undertaking the Representing Genes study, and experimental philosophy of biology in general. X-phi of biology is a potentially important part of what Paul Griffiths has called the 'biohumanities'.[1] The concept of 'biohumanities' is a vision of the relationship between the humanities (including philosophy of science), biology and society (Stotz & Griffiths, 2008). In this vision, the humanities create knowledge *about biology*. Both the history of genetics and philosophical work on the concept of the gene of the sort described in this paper enrich our understanding *of genetics itself*. Contrast this to the vision implicit in the idea of Ethical, Legal and Social Implication ('ELSI') research, in which biologists provide the facts and humanists and social sciences work out their implications for society, or to one traditional vision of history and philosophy of science, in which studies of particular sciences are data for conclusions about the nature of science. In the biohumanities vision, good history and sound philosophy may provide resources for addressing 'ELSI' issues, but that is not their primary aim. Likewise, we may learn something about the nature of science from our work, but it is enough if we learn something about genetics, ecology or whatever other bioscience we study.

We can discern four aims of research in the 'biohumanities'. First, Biohumanities understands itself as a critical enterprise. Constructive 'science criticism' (Pigliucci & Kaplan, 2006, p. 8) stands back from the urgencies of actual scientific research to

E-mail address: karola.stotz@gmail.com
[1] http://paul.representinggenes.org/biohum_home.html.

0039-3681/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.shpsa.2009.03.008

reflect on the strengths and weaknesses of current approaches. Thus, as C. Kenneth Waters has remarked, the aims of conceptual analysis in the philosophy of science include:

> to articulate scientific concepts in ways that help reveal epistemic virtues and limitations of particular sciences. This means an analysis of the gene concept(s) should help clarify the explanatory power and limitations of gene-based explanations, and should help account for the investigative utility and biases of gene-centered sciences. (Waters, 2004, p. 29)

Secondly, science criticism, without necessarily questioning any specific findings of past research, can simply be the 'turning-over of stones that had hitherto held their ground' (Moss, 2006, p. 523). History of science points out the 'roads not taken' in science. Without calling into question the data that was gathered, it points out that other sorts of data might have been gathered, or that the data that was actually gathered could have been interpreted in different ways. Philosophy of science adds to this enterprise by critically analyzing the chains of reasoning that connect specific scientific findings to claims about the broader significance of those findings. This can lead to changes in interpretation that can potentially motivate biologists to reinterpret earlier scientific findings and to prioritize different questions for future research. Ideally, history and philosophy of biology can articulate *alternative visions* of biology.

Thirdly, good history and philosophy of biology can contribute to the creation of 'critical science communication' through promoting a critical understanding of scientific claims in the sense just outlined and communicating to a wider audience not merely 'what has been discovered', but something of the complexity of the scientific process and the contestability of its findings. To be useful, critical work of the sort described in the last two paragraphs must be 'bioliterate', something that might be roughly defined as engaging with the science at the same level as practitioners, rather than via popular representations. But the broad 'visions' of science in which it results can be expressed in a non-technical way, and can thus make a major contribution to the public understanding of science.

Fourthly, and most generally, biohumanities is concerned with *understanding biology*. Although it should by now be clear that I think the biohumanities are of potential value to both biology and society, this is not the only possible justification for biohumanities research. Science is fascinating and important, and it is worth understanding even if understanding it does not make it work any better, just as evolution is worth understanding whether or not doing so contributes to crop improvement or drug development. As in the sciences themselves, it is hard to imagine researchers doing their best work without an intrinsic interest in the material they study.

It may be worth reiterating at this point that professional historians and philosopher do not have a monopoly on any of the activities described in this section. Anyone acquainted with the literature in the history and philosophy of biology knows that biologists have made very substantial contributions to both. But when making these contributions, they are not doing biology, but history or philosophy, or 'science criticism'.[2]

The following sections of the paper will give more substance to these remarks with the specific example of the Representing Genes Project conducted by myself and my collaborators. Section 2 outlines the project. Section 3 uses the project to introduce the ideas of scientific concepts as 'tools' for research and the experimental philosopher of science as a 'conceptual ecologist'. In accordance with these two metaphors, the aim of the Representing Genes Project was to reveal complementary and competing gene concepts and locate them in different areas of biological research. Our conclusions about gene concepts and their epistemic niches are presented in Section 4. Finally, in Section 5, we argue for a new vision of the gene suited to postgenomic biology.

2. Representing genes

The gene's conceptual variation is the most salient feature of its 100-year history. Philosophical analyses have attempted to both *describe* different concepts of the gene in use by different biologists and at different times, and to *prescribe* better ones (Waters, 2004). Many philosophers and historians have analyzed this variation in terms of a tension between two aspects of the gene, aspects that have been characterized in several, not necessarily consistent ways:

- An abstract unit of inheritance versus a concrete entity (Falk, 2000)
- A functional or top down versus a structural or bottom up approach to the gene (Gilbert, 2000).
- 'Gene P' (for phenotype, predictor or preformation) versus 'Gene D' (for developmental resource) (Moss, 2003), or 'pop gene' (of evolutionary genetics) versus 'dev gene' (of developmental genetics) (Gilbert, 2003)

One general feature that might help to unify these competing characterizations would be to distinguish between a statistical and a mechanistic relationship between the gene and phenotypic traits. Others have approached the variation in the gene concept, not by introducing distinctions, but by unifying the variants under a more general concept, such as schematic gene which yields different specific genes along a continuum from proximal to distal relationships to gene products (Waters, 1994, 2000). In my own work my collaborators and I have come to embrace a version of the dichotomy between an abstract, statistical gene and a concrete, mechanistic gene, but have felt the need to introduce a further distinction between a simple stereotype or 'consensus gene' versus a gene which embraces the complex and messy reality of the relationship between genome structure and genome function (Stotz et al., 2006; Griffiths & Stotz, 2006).

The Representing Genes Project[3] was an attempt to empirically assess the impact of the ongoing molecular genomics revolution on concepts of the gene (Stotz & Griffiths, 2004; Stotz et al., 2006). The survey instrument developed for the Representing Genes study was designed to explore several of the issues posed by the existence of alternative gene concepts.

Part 1 of the questionnaire was inspired by a pressing problem posed by the move into the genome-sequencing era, namely which principles to use in order to functionally annotate DNA sequences. The actual practice of genome annotation inspired us to design a simple, annotation-like task to investigate the criteria that lead biologists to annotate a particular DNA sequence as either one gene with several gene products or several genes with a single functional product. This 'simplified annotation' task used graphical representations and descriptions of real DNA transcription events in eukaryotic genomes which challenge various aspects of the classical molecular gene concept. The cases were chosen to allow pair-wise comparisons highlighting differences in the criteria that may influence biologists' judgments on this issue during annotation. The examples illustrate the flexibility, variability and complexity of 'genome expression' that complicate defining what

[2] A term we have in fact taken from the joint work of Massimo Pigliucci, a biologist, and Jonathan Kaplan, a philosopher.

[3] *Representing genes: Testing competing philosophical analyses of the gene concept in contemporary molecular biology* (2006). http://representinggenes.org.

genes are. Since common definitions of the gene are insufficient for making this decision, the simplified annotation task is designed to reveal the implicit criteria biologists draw upon in this judgment.

To give a flavor of the work, I will sketch two of the cases, and the conclusion we drew from them. One phenomenon not anticipated by the classical molecular conception of the gene is 'antisense transcription', in which the two strands of the DNA double-helix each contain their own genes, running in opposite directions. Another is '*trans*-splicing', in which messenger RNAs transcribed from different parts of the genome are spliced together at a later stage to make a single molecule. This means that a single polypeptide may be derived from several loci in the genome. These phenomena come together in two of our cases which were based on events at the Drosophila locus *mod(mdg4)*. This stretch of DNA contains a fairly conventional-looking gene from which the larger part of a messenger RNA molecule is made by the usual cutting and pasting. The final section of the molecule, however, comes in several variations, which are made from various short sections of DNA located some distance away on the chromosome and *trans*-spliced to the larger section. Some of these are on the same strand of the DNA double helix as the main part of the molecule. Others are on the opposite strand of the double helix, in which case they run in the opposite direction along the DNA. The mechanisms that make the smaller part of the molecule and join it to the larger part are identical in both cases. We find, however, that while biologists predominantly regard the first case as a single gene, they predominantly regard the second case as several genes cooperating to make a single product. What explains this difference in their responses?

In recognition of the difficulties of the classical molecular gene concept, it has been suggested that working biologists employ a *consensus gene* concept based on a collection of flexibly applied features of well established genes. On this proposal, there is something similar to a threshold effect for considering a stretch of DNA to be a gene if it has 'enough' of these features, such as an open reading frame, a consensus core promoter sequence (the TATA box), or the existence of RNA transcripts. The originator of this view, Thomas Fogle, has argued that by combining structural and functional features into a single stereotype, the consensus gene concept hides both the diversity of DNA sequences that can perform the same function and the diverse functions of particular DNA sequences (Fogle, 2000). In other words, the consensus gene concept inherently distracts from conceptually problematic cases. It means that when biologists come to annotate a locus at which something unusual is happening, they look for ways to divide the sequence into several different genes, none of which diverges too far from the stereotype. We suggest that just this has happened in our two cases based on the *(mod)mdg4* locus. Although the cases in which the alternative terminal exon is *trans*-spliced from the same strand are not significantly different in terms of molecular mechanisms from those in which it is *trans*-spliced from the antisense strand, the idea of a single gene with parts running in opposite directions—'head to head'—is just too unlike the prototypical gene pictured in textbooks, and so in this case they are seen as separate genes. Whether something is one gene or two is thus as much a psychological as a biological matter.

Another part of the survey instrument set out to investigate whether and when, as Lenny Moss has argued (Moss, 2003), investigators either start with the conception of a theoretical gene determinately connected to a particular phenotype, or of a concrete molecular gene with a specific sequence and the template capacity to code for many products, depending on how it is transcribed and how its initial product is later processed. We argued that these different starting points would affect how investigators set out to unravel the complex relationship between genes and other molecular factors with the phenotype. Hence the second task asked subjects to assess the value of different research strategies for investigating complex diseases. For each disease we offered four strategies, designed to run along a continuum from focusing on the statistical relationship between gene and phene to entirely giving up on such a relationship in favor of analyzing content-dependent causal pathways between the two. We looked for differences in which strategies were favored by biologists from different backgrounds, and also at whether the choice of strategies changed between human versus animal disease, and for physiological versus psychological disease.

In this section, I have given a brief description of some experimental philosophy of biology. In the next section I turn to the issue of what philosophers can learn from this kind of work.

3. The philosopher as a scientist

One motivation for the Representing Genes study was to transcend the limitations of traditional conceptual analysis. There is a tradeoff between the intimate knowledge of part of the science through interaction with particular scientists and the bias of your interpretation of the whole field. Perhaps as a result, philosophers typically produce competing analyses of scientific concepts, and traditional conceptual analysis too often ends with the 'dull thud of conflicting intuitions'. Such problems have produced increasing interest in bringing a new set of tools to bear. Experimental philosophy of science has the capacity to assess these competing analyses against data and to avoid biases introduced by working with a single subdiscipline or a single school of thought.

Such a philosophy 'in the trenches' is also in a privileged position to provide the bridge between philosophy and science. The 'trench' of the experimental philosopher does not demarcate the line between the humanities and science in the science war, but the empirical frontline in the fight for real knowledge where philosophy and science unite. At least part of philosophy of science has abandoned the idea that its job is to enforce rigor and precision within science through the fixation of scientific meaning. It has been argued that slippage of meaning was essential to the rapid progress of genetics (Rheinberger, 2000). Equally gone is Paul Feyerabend's conceptual anarchism, in which the history of science is little more than a series of changes in the fashionable topics of scientific discussion (Feyerabend, 1975). In place of these two models we have come to appreciate that conceptual change in science is rationally motivated by what scientists are trying to achieve, by their accumulated experience of how to achieve it, and by changes in what they are trying to achieve. Empirical science is a powerhouse of conceptual innovation because scientists use and reuse their terminology in a truly 'exuberant' way (Rheinberger, 2000). The gene concept is a case in point: despite its ever-changing definition, the gene remains on the laboratory bench after a whole century because it has proved a flexible tool.[4] This only makes sense if we think of concepts as tool of research, as ways of classifying the experience shaped by experimentalists to meet their specific needs. Necessarily these tools get reshaped as the scientists' needs change.

In the study of conceptual diversification, the history of genetics provides a 'conceptual phylogeny' of the gene and the Representing Genes Project can be seen as an attempt to determine some of the

[4] But see Moss (2006) with his critique of using the 'gene' as placeholder for a full explanation of life.

'ecological' pressures that have caused the gene concept to diversify into different 'epistemic niches'. The next section presents an attempt to describe the phylogeny and ecology of the former and current use of the gene concepts.

4. 'Conceptual speciation events' and the 'epistemic niche'[5]

The gene was originally defined in the light of the hybridization techniques available to early geneticists. In the absence of any knowledge about the molecular basis of genetics this early 'instrumental gene' was a hypothetical entity, an intervening variable between the phenotypes of the parental generation and the distribution of phenotypes in following generations. As new techniques became available and new questions about the structural nature of the gene pressing, the gene was redefined. However, just as old techniques can survive alongside newer ones, old concepts can remain the best tool for the work for which they were originally designed. For example, when a medical geneticist is seeking the 'genes for' a disorder she is looking for traditional Mendelian genes—sections of chromosome whose pattern of inheritance explains the phenotypic differences observed in patients. Translated into molecular terms these sections may turn out not to be molecular genes. Some abnormalities in human limb development, for example, have been tracked down to mutations in a gene on chromosome 7. But recent research suggests that the gene in which the mutation is located plays no role in the development of these abnormalities (Lettice et al., 2002). Instead, embedded in that gene is a sequence which acts to regulate the use of the gene 'sonic hedgehog', about one million DNA nucleotides away on the same chromosome, which is involved in the relevant aspects of limb development. Nothing is gone wrong in either piece of research. It is simply that the molecular gene concept is not a good tool for some kinds of research. The instrumental, Mendelian gene remains the best tool in fields like medical genetics and population genetics. So while a particular scientific concept reflects the scientific knowledge at a point in time, this alone cannot explain the parallel use of several different concepts. For a full understanding of that phenomenon we need to see scientific concepts as tools for research, as much as glassware, microscopes or scales.

In the 1960s molecular biologists believed they had arrived at a single molecular concept of the gene, which united the structural and functional aspects of the gene. The molecular gene is a *structure* in the DNA whose *function* is to specify the linear order of elements in a gene product (RNA or polypeptide). This is reflected in the Central Dogma of Molecular Genetics, which claims that the genetic information, the linear sequence of nucleic acid bases, specifies the linear order of the gene product, with no feedback mechanisms allowed (Crick, 1958, 1970). In the light of today's knowledge about the ways in which a limited number of DNA sequences is used to create a vastly greater 'transcriptome' of gene products, the sequences we count to arrive at the claim that there are about 21,000 human or 14,000 *Drosophila* protein coding genes are best regarded as *stereotypical* genes—sequences that fit a stereotype of how DNA plays the gene-role. The cases that inspire the stereotype are the simple cases of bacterial transcription and translation that were used to derive our basic understanding of molecular genetics in the 1960s. The prototype is undermined by heavily edited mRNA transcripts derived from 'cryptogenes' or by cases of *trans*-splicing, in which the linear order of the product is no longer mirrored in the linear order of nucleotides in the DNA.

How does a scientist today decide where one gene starts and another stops? Since one gene can code for many different products when expressed in different ways or in cooperation with other sequences, there is no principled answer. The tip of the iceberg of the complexity of gene expression is the common process of alternative splicing.[6] It was the first mechanism detected to seriously undermine the one gene–one polypeptide hypothesis enshrined in the classical molecular gene concept. However, the widespread convention of molecular geneticists is to define alternatively spliced genes as *one* gene with *multiple* products. At first surprising, the underlying rationale of the molecular gene concept actually explains this convention: A molecular gene is defined as the linear image of a gene product in the DNA (Waters, 1994, 2000). Accordingly, if, in the scientist's judgment, the many different products of a single DNA sequence are sufficiently similar, for example protein isoforms with many shared functional subunits, then they are produced from one gene. If, however, they are *sufficiently different* from one another, for example through the process of frame-shifting, then they are the products of *two* overlapping genes. As we discussed in Section 2, a key aim of genome annotation is to find a way of segmenting complex eukaryote genomes into sequences that look reasonably like the prototype of a gene (Fogle, 2000).

5. A postgenomic gene concept?

Just as finer work may require more specialized tools, it may be that molecular geneticists are now confronting problems for which the classical molecular gene concept no longer proves useful. For instance, scientists today want to understand how regulated genome expression leads from a ridiculously small number of genes to the *explosion* of gene products that create and maintain higher organism, especially humans. This will likely require a more modest gene concept in which the structural and functional aspects of the molecular gene are dissociated again. In reality the way in which the DNA contains the image of its product is often akin to the way in which Picasso's cubist paintings contain a fragmented and distorted image of his models. The best trick of the genome, however, the cubists never invented, namely how a single sequence (brushstroke) can be part of many genes (paintings). We now know that complex forms of transcriptional and post-transcriptional processing, at least in eukaryote genomes, are 'business as usual'. Beside the well known process of alternative splicing, more recently it has been discovered that through *trans*-splicing coding sequences can be pasted together in a different order, repeated or even pasted in backwards. Thus, as well as getting many products from a single piece of DNA, several pieces of DNA can be used to make a single gene product. These pieces may even be located on different chromosomes. Sequences can be transcribed or translated in different reading frames, or be edited through the insertion, deletion or substitution of nucleotides (Stotz, 2006a,b).

The postgenomic gene concept (Griffiths & Stotz, 2006, 2007), rather than covering up any unwanted messiness, welcomes and embraces these complexities in the relationship between DNA and its products as new opportunities to relate a mere 21,000 genes to all the complexities of human development and functioning. Thanks to a large variety of specific complexes of interacting regulatory molecules (*cis*-acting sequences in the genome, *trans*-acting factors of gene products, metabolites and other environmental signals) DNA is used in highly time- and tissue-specific ways. Regulated recruitment and combinatorial control

[5] For a more detailed explanation of the points in this and the next sections, see Stotz et al. (2006), Stotz (2006b).

[6] The majority of genes in higher organisms are alternatively spliced and the current star example of alternative splicing is *Dscam* (Drosophila Cell Adhesion Molecule), a gene which may produce up to 38,016 different forms of the DSCAM protein.

of these regulatory molecules is the mechanism of choice of most organism to control gene expression (Ptashne & Gann, 2002). Lenny Moss has described these as 'ad hoc committees' of regulatory molecules whose particular 'membership' reflects the contingent history of the cell up to that time, including the history of the cell's transactions with its environment (Moss, 2003). This metaphor is designed to embody the new biology of genome regulation in the same way that the metaphor of a genetic program written in the DNA embodied the biology of the 1960s.

In this vision the 'gene' is relieved of its unrealistic and mystical status as the sole embodiment of life with a (unsurprisingly, not very well understood) propensity to 'work on its own behalf' (Kauffman, 2000; Moss, 2006). Instead, genes become prosaic ways to classify the template capacity of certain parts of the genome, a capacity that must be interpreted through a process of gene expression to yield any determinate result. Because of this limited and very context-dependent capacity, the gene is also stripped of its place as the sole unit of inheritance. Predictable expression patterns of parts of the genome are ensured by the reliable reproduction of a developmental niche that regulates the same expression patterns. Inheritance is not embodied in mystical preformations of the phenotype but in the reproduction of the necessary factors of development that will self-organize to reproduce a similar developmental life cycle. Life is not situated in genes but the particular organization of biomolecules that enables the system to maintain itself by reconstituting its own components from the template capacity in the genome, constructing the environmental factors necessary for this to occur, and ultimately reproducing copies of itself.

6. Conclusion

This paper describes empirical/experimental studies in the philosophy of biology in general as part of the 'biohumanities'. This field comprises four different but related aims: constructive science criticism, creating alternative visions of biology, critical science communication, and, simply, understanding biology as an object of natural knowledge in its own right. In Section 3 I outlined how experimental philosophy methods can contribute to this kind of research. Section 4 and 5 demonstrated the critical potential of this research when directed at current molecular biology.

Acknowledgement

This material presented in this paper is based upon work supported by the National Science Foundation under Grants #0217567 and #0323496, awarded to the author and Paul Griffiths.

References

Crick, F. H. C. (1958). On protein synthesis. *Symposia of the Society for Experimental Biology, 12*, 138–163.
Crick, F. H. C. (1970). Central dogma of molecular biology. *Nature, 227*, 561–563.
Falk, R. (2000). The gene: A concept in tension. In P. Beurton, R. Falk, & H.-J. Rheinberger (Eds.), *The concept of the gene in development and evolution* (pp. 317–348). Cambridge: Cambridge University Press.
Feyerabend, P. (1975). *Against method*. London: Verso.
Fogle, T. (2000). The dissolution of protein coding genes in molecular biology. In P. Beurton, R. Falk, & H.-J. Rheinberger (Eds.), *The concept of the gene in development and evolution* (pp. 3–25). Cambridge: Cambridge University Press.
Gilbert, S. C. (2000). Genes classical and genes developmental: The different uses of genes in evolutionary syntheses. In P. Beurton, R. Falk, & H.-J. Rheinberger (Eds.), *The concept of the gene in development and evolution* (pp. 178–192). Cambridge: Cambridge University Press.
Gilbert, S. F. (2003). Evo-devo, devo-evo, and devgen-popgen. *Biology and Philosophy, 18*(2), 347–352.
Griffiths, P. E., & Stotz, K. (2006). Genes in the postgenomic era. *Theoretical Medicine and Bioethics, 27*(6), 499–521.
Griffiths, P. E., & Stotz, K. (2007). Gene. In D. Hull, & M. Ruse (Eds.), *Cambridge companion for the philosophy of biology* (pp. 85–102). Cambridge: Cambridge University Press.
Griffiths, P. E., & Stotz, K. (2008). Experimental philosophy of science. *Philosophy Compass, 3*(3), 507–521.
Kauffman, S. A. (2000). *Investigations*. Oxford: Oxford University Press.
Lettice, L. A., Horikoshi, T., Heaney, S. J. H., Baren, M. J. van, Linde, H. C. van der, Breedveld, G. J., et al. (2002). Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proceedings of the National Academy of Sciences, 99*(11), 7548–7553.
Moss, L. (2003). *What genes can't do*. Cambridge, MA: MIT Press.
Moss, L. (2006). The question of questions: What is a gene? Comments on Rolston and Griffiths & Stotz. *Theoretical Medicine and Bioethics, 27*(6), 523–534.
Pigliucci, M., & Kaplan, J. (2006). *Making sense of evolution: The conceptual foundations of evolutionary biology*. Chicago: University of Chicago Press.
Ptashne, M., & Gann, A. (2002). *Genes and signals*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
Rheinberger, H.-J. (2000). Gene concepts: Fragments from the perspective of molecular biology. In P. J. Beurton, R. Falk, & H.-J. Rheinberger (Eds.), *The concept of the gene in development and evolution* (pp. 219–239). Cambridge: Cambridge University Press.
Stotz, K. (2006a). Molecular epigenesis: Distributed specificity as a break in the Central Dogma. *History and Philosophy of the Life Sciences, 28*(4), 527–544.
Stotz, K. (2006b). With genes like that, who needs an environment? Postgenomics' argument for the ontogeny of information. *Philosophy of Science, 73*(5), 905–917.
Stotz, K., Bostanci, A., & Griffiths, P. E. (2006). Tracking the shift to 'post-genomics'. *Community Genetics, 9*(3), 190–196.
Stotz, K., & Griffiths, P. E. (2004). Genes: Philosophical analyses put to the test. *History and Philosophy of the Life Sciences, 26*(Special issue *Genes, genomes and genetic elements*), 5–28.
Stotz, K., & Griffiths, P. E. (2008). Biohumanities: Rethinking the relationship between biosciences, philosophy and history of science, and society. *Quarterly Review of Biology, 83*(1), 37-45.
Stotz, K., Griffiths, P. E., & Knight, R. D. (2004). How scientists conceptualize genes: An empirical study. *Studies in History and Philosophy of Biological and Biomedical Sciences, 35*(4), 647–673.
Waters, C. K. (1994). Genes made molecular. *Philosophy of Science, 61*, 163–185.
Waters, C. K. (2000). Molecules made biological. *Revue Internationale de Philosophie, 4*(214), 539–564.
Waters, C. K. (2004). What concept analysis should be (and why competing philosophical analyses of gene concepts cannot be tested by polling scientists). *History and Philosophy of the Life Sciences, 26*, 29–58.

Studies in History and Philosophy of Science 40 (2009) 238–242

Contents lists available at ScienceDirect

Studies in History and Philosophy of Science

journal homepage: www.elsevier.com/locate/shpsa

Folk judgments of causation

Joshua Knobe

Department of Philosophy, CB # 3125, Caldwell Hall, UNC/Chapel Hill, Chapel Hill, NC 27599-3125, USA

ARTICLE INFO

Keywords:
Causation
Causal cognition
Experimental philosophy
Cognitive science

ABSTRACT

Experimental studies suggest that people's ordinary causal judgments are affected not only by statistical considerations but also by *moral* considerations. One way to explain these results would be to construct a model according to which people are trying to make a purely statistical judgment but moral considerations somehow distort their intuitions. The present paper offers an alternative perspective. Specifically, the author proposes a model according to which the very same underlying mechanism accounts for the influence of both statistical and moral considerations. On this model, it appears that ordinary causal judgments are quite different from the sorts of judgments one might find in the systematic sciences.

1. Introduction

When scientists are trying to uncover the causes of a given outcome, they often make use of statistical information. Thus, if scientists wanted to know whether there was a causal relationship between attending philosophy lectures and learning philosophy, they might randomly assign students either to attend or not attend certain lectures and then check to see whether those who attended the lectures ended up learning more philosophy than those who did not.

A question now arises as to how ordinary people—people who have no formal training in the sciences—typically uncover the causes of particular outcomes. One popular answer to this question is that ordinary people use more or less the same techniques that scientists do (e.g. Kelley, 1967; Woodward, 2003; Gopnik et al., 2004). Of course, ordinary people do not actually write out equations or use precise statistical methods, but one might nonetheless claim that they uncover causal relations by looking in a more informal way at statistical information and that they actually look for pretty much the same sorts of statistical information that scientists do. At least within social psychology, this view is associated with the slogan *The person as scientist*.

Although this 'person as scientist' theory remains the dominant view both in philosophy and in psychology, recent years have seen the emergence of a new research program whose results point in a radically different direction. Research within this program has shown that people's causal judgments can sometimes be influenced by their *moral* judgments. In other words, when people are wondering about the causal relationships between events, their conclusions can be influenced by their beliefs as to whether those events are morally good or morally bad. At least on the surface, these results seem to serve as a challenge to the view that people assess causation by making something like a scientific judgment.

Yet researchers have not generally reacted by questioning prior assumptions about the nature of people's causal judgments. Instead, the usual view is that people truly are trying to make a purely statistical judgment but that certain processes are leading to 'distortions' or 'biases' in these judgments. The basic idea behind this view can be captured by the slogan *The person as bumbling scientist*. In essence, what it says is that people are engaged in an attempt to make scientific judgments but that they are messing up somehow and thereby allowing moral considerations to influence their intuitions.

My aim here is to offer an alternative hypothesis. I posit a single underlying mechanism that explains both the impact of statistical considerations and the impact of moral considerations. The claim, then, is that we should abandon the idea that causal judgments are fundamentally statistical and that the influence of moral considerations constitutes some sort of 'bias' or 'distortion'. In its place, we can adopt a theory according to which moral considerations truly

E-mail address: Knobe@email.unc.edu

0039-3681/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.shpsa.2009.03.009

do play a role in the fundamental mechanisms underlying causal judgments.

2. The role of statistical considerations

George writes out what he regards as a profound and original idea and sends it off to an academic journal ... But there is a bad outcome. The paper gets a negative review and is therefore rejected. In a situation like this, one can easily imagine a person wondering what exactly caused the bad outcome. Was it something about the paper itself? Or something about the reviewer? Or a combination of the two?

Of course, there is a fairly obvious sense in which both facts about the reviewer and facts about the paper stand in some sort of causal relation to the problem that results, and one could well imagine that people might be content simply to trace out these various causal relations and understand how each of them works. Yet it appears that people do not actually proceed in that way. Instead, they seem to select certain particular factors and refer to those alone as 'causes', while classifying all the others as mere 'background conditions' or 'enabling factors'. This process has come to be known as *causal selection*, and it plays an important role in folk judgments of causation.

A long tradition of research within social psychology has shown that the process of causal selection is sensitive in systematic ways to *statistical* considerations. So, for example, consider the following two cases:

> (Case 1) George sends his paper to a number of different journals and conferences, and they all reject it. Meanwhile, the reviewer accepts a number of other papers written by other authors.
>
> (Case 2) George sends his paper to a number of different journals and conferences, and they all accept it. Meanwhile, the reviewer rejects every single paper he is given.

Research shows that people's causal judgments about George's paper will depend on which of these two cases is the actual one (e.g. McArthur, 1972; Hilton & Slugoski, 1986). People will tend to say that the problem was caused by George's paper in Case 1, whereas they will tend to say that the problem was caused by the reviewer in Case 2.

It is really quite a striking fact that people respond in this way. After all, even if the reviewer was disposed to reject 99.99% of philosophy papers, one could still say that the bad outcome actually was caused by something about the paper—namely, the fact that it wasn't one of the .01% of papers that the reviewer would be inclined to accept. Yet the available research shows that people tend not to respond in that way. Instead, when faced with a case like this one, they say that the bad outcome *was* caused by something about the reviewer but *wasn't* caused by something about the paper. What we want to understand now is why exactly people take statistical considerations into account in this way.

The usual view within social psychology is that people's judgments in such cases are more or less analogous to the judgments one might make in the course of a scientific inquiry. Suppose that a scientist was trying to figure out whether the fates of academic papers were mostly due to something about the papers themselves or whether they were mostly due to something about the individual reviewers. The first step would probably be to conduct an *analysis of variance* (ANOVA). One could give a lot of different papers to a lot of different reviewers and try to figure out what percentage of the total variance was explained by facts about the papers and what percentage was explained by facts about the reviewers. (As it happens, this experiment has actually been conducted. The answer is that a substantial percentage of the variance is explained by facts about the reviewers and almost none is explained by facts about the papers; Cole et al., 1981; Blackburn & Hakel, 2006).

Of course, no one suggests that ordinary people make causal judgments by using precisely the same mathematical procedures that scientists use when calculating an ANOVA, but many researchers have suggested that we can think of the process underlying ordinary causal judgments as being *similar* in certain ways to the calculation of a full-fledged ANOVA (Kelley, 1967; Försterling, 1989).[1] To a first approximation, the claim is that people tend to attribute outcomes to whichever factor they think explains the greatest percentage of the variance. If they think that most of the variance is explained by facts about reviewers and almost none is explained by facts about the papers themselves, they will tend to say that the bad outcome was caused not by anything about the paper but solely by something about the reviewer.

3. The role of moral considerations

But it seems that things are not quite so simple. As a number of studies have shown, people's causal judgments can be influenced not only by statistical considerations but also by *moral* considerations (Alicke, 1992; Solan & Darley, 2001; Cushman, 2006; Knobe & Fraser, 2008). That is, when people are wondering whether *x* caused *y*, their judgments depend in part on whether they believe that *x* itself is morally good or morally bad.

Perhaps the best way of conveying the basic issues here is to give a simple example. In one recent experiment, subjects were given the following vignette:

> The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take the pens, but faculty members are supposed to buy their own.
>
> The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist has repeatedly emailed them reminders that only administrative assistants are allowed to take the pens.
>
> On Monday morning, one of the administrative assistants encounters Professor Smith walking past the receptionist's desk. Both take pens. Later that day, the receptionist needs to take an important message ... but she has a problem. There are no pens left on her desk. (Knobe & Fraser, 2008, p. 443)

From a statistical perspective, the behavior of the faculty member and the behavior of the administrative assistant seem more or less the same. Both agents performed behaviors of quite ordinary sorts, and each behavior was related to the ultimate outcome in the same way. Yet subjects did not treat these two behaviors alike. They judged that the faculty member *did* cause the problem but that the administrative assistant *did not* cause the problem.

What we see here, apparently, is an impact of moral considerations on causal judgments. It seems that people classify the faculty member's behavior as *wrong* and that this classification then influences their judgments about whether his behavior caused the problem. The key question now is why exactly this effect arises.

One obvious way to react here would be to conclude that results like this one call into question the 'person as scientist' theory. Hence, one might say: 'If people really were behaving like

[1] Recent years have seen the emergence of more complex statistical frameworks (e.g. Pearl, 2000; Spirtes et al., 2000), but unlike the ANOVA model, these frameworks do not purport to solve the problem of causal selection. It would be possible, then, to develop a model according to which judgments about the underlying causal structure were purely scientific but the process of causal selection then introduced an additional element that went beyond questions that could be solved using scientific methods alone (see, for example, Hitchcock, 2007).

scientists, they would not allow moral considerations to influence their judgments. So the available evidence now suggests that people actually aren't behaving like scientists but are instead engaged in some very different sort of inquiry'. But that has not been the usual reaction thus far. Instead, the usual reaction has been to suggest that there is a sense in which the original theory was actually right and it is the experimental subjects themselves who are wrong. That is to say, researchers react to these results by suggesting that subjects truly are trying to engage in a purely statistical inquiry but that some further process is *interfering* here and allowing moral judgments to shape people's responses.

One way to address this issue would be to look in detail at a number of specific hypotheses about precisely how people's moral judgments could interfere with the proper operation of their causal reasoning. Some researchers have suggested that the effects here are due to a motivational bias (Alicke, 1992, in press); others have suggested that the effects are due to pragmatics (Driver, 2008a, 2008b). We could examine each of these hypotheses in turn, looking for confirming or disconfirming evidence. This strikes me as an extremely valuable project (and one I have attempted elsewhere; Hitchcock & Knobe, in press; Knobe & Fraser, 2008), but I will not be pursuing it here.

Instead, I want to proceed by offering a positive hypothesis about the underlying psychological mechanisms that generate these effects. On this hypothesis, there is no sense in which statistical considerations truly do play a role in the fundamental competence while moral considerations are merely a 'distortion' or 'bias'. On the contrary, the hypothesis suggests that *the very same process* explains the use of both statistical and moral considerations.

4. The classification of counterfactuals

In particular, I want to suggest that both of these effects can be explained in terms of a quite general theory about people's capacity for counterfactual reasoning. Hence, the suggestion will be that statistical and moral considerations end up having an impact on people's causal judgments because they have an impact on the way people reason about counterfactuals in general.

The first thing to note here is that our capacity for counterfactual reasoning not only allows us to distinguish between counterfactuals that are true and those that are false but also to distinguish between counterfactuals that are worth considering and those that we would do better simply to ignore. Thus, suppose that I end up getting a bad grade on an important test. I might immediately begin considering certain counterfactuals: 'What if I had studied harder?', 'What if I had chosen not to go to that party last night?' And many other similar thoughts might fill my mind. But there are also counterfactuals that I would regard as silly and not worth considering. I would never think: 'What if the teacher had been struck by a meteor before she finished grading my exam?'

It is an important fact about human cognition that we are able to focus in this way on the relevant counterfactuals and ignore the irrelevant ones. Counterfactual thinking can help us to solve many practical problems, but if we ended up reflecting in detail on any counterfactual that came to mind, we would never be able to get anything done. Ideally, then, our minds would somehow manage to focus in on the relevant counterfactuals and suppress all further thoughts about the irrelevant ones. It seems that human beings actually are surprisingly good at accomplishing this task.

There has been a great deal of experimental research in psychology on people's capacity for counterfactual reasoning, and a considerable amount is now known about the precise conditions under which people do and do not consider various counterfactuals. Here, however, we can proceed by borrowing just three basic principles from this vast literature:

- The first principle says that people are inclined to consider counterfactuals in which events of statistically *unusual* types are replaced by events of statistically *usual* types (Kahneman & Tversky, 1982; Roese, 1997). Example: if a student hands in her paper on a roll of toilet paper, we will be inclined to think about what would have happened if she had handed it in on computer paper instead.
- The second principle says that people are inclined to consider counterfactuals in which *bad* events are replaced by *good* events (Read, 1985; Kahneman & Miller, 1986). Example: if a committee makes a bad decision, we will be inclined to think about what would have happened if it had instead made a good decision.
- The third principle establishes a kind of default. It says that, unless there is some specific reason to think about a given counterfactual, people will be inclined to classify it as irrelevant and not give it any further thought.[2]

In articulating this third principle, I depart in certain respects from the traditional way of thinking about these phenomena within psychology. The traditional view is that certain types of counterfactuals just never occur to people in the first place. The hypothesis being offered here is that there is actually something more complex going on. It is not just that certain counterfactuals never occur to people; it is that these counterfactuals are actually *classified as irrelevant*. To see the force of this claim, consider what might happen if we actively intervened in a student's life and forced him to consider what would have happened if his teacher had been struck by a meteor. (One way to arrange such an intervention would be to ask the student to write a detailed essay on the topic.) After our intervention was complete, the student would have very vivid and definite views about precisely what would have occurred under the specified counterfactual conditions. Nonetheless, the hypothesis is that the student would classify the whole counterfactual as irrelevant and that his beliefs about it would therefore have little impact on any further cognitive processes he might undergo. Hence, if we later asked him to think about how he might avoid getting bad grades in the future, he would not begin by thinking: 'Well, I would have avoided this bad grade if my teacher had only been struck by a meteor ...' And similarly for any other aspect of cognition. No matter where we look, we will find a certain resistance to considering counterfactuals that have been classified as irrelevant.

Ultimately, the hope is that we can draw on these very general claims about people's use of counterfactuals to explain the puzzling experimental results concerning people's judgments of causation. It is important to emphasize, however, that our claims about people's use of counterfactuals do not themselves depend on evidence from studies of causal reasoning. Instead, these claims are based on *independent* evidence. (The experimental studies look directly at people's counterfactual reasoning, not at their judgments of causation.) Our overall theory can therefore draw on two different sources of experimental evidence. First we look at studies of counterfactual reasoning and thereby construct a general theory about how people use counterfactuals; then we take this general theory and try to use it to explain the results of experiments about people's causal judgments.

[2] Ultimately, it might be possible to unify these three principles in the single claim that people regard a counterfactual as relevant to the extent that it replaces *abnormal* things with *normal* ones (Hitchcock & Knobe, 2008). The key assumption would be that people have a notion of 'normality' that includes both statistical and moral elements.

5. Explaining the effects

To get from this general theory of counterfactuals to specific predictions about causal judgments, we need to introduce an additional assumption, namely, the assumption that people arrive at causal judgments by making use of counterfactuals.

Fortunately, we have good reason to believe that this assumption is correct. A wide variety of theories of causal judgment suggest that these judgments actually do rely on counterfactuals in one way or another (e.g. Lewis, 2000; Collins et al., 2001; Woodward, 2003; Hitchcock, 2008). These theories differ from each other in a number of important respects, but those differences will not be relevant here. Instead, we will simply be relying on the basic claim that people make judgments about whether a given event caused an outcome by considering counterfactuals in which that event does not occur.

Armed with this assumption, we can now reexamine the experimental results concerning people's causal judgments. The aim will be to show that it is possible to explain the patterns we observe in these judgments by drawing on a general theory of counterfactual reasoning.

First, consider the role of statistical considerations. Our example here was the paper that had been submitted for review at an academic journal. The paper, let us suppose, is good enough that it would normally be accepted, but it has been sent to a reviewer who has a tendency to reject almost every manuscript he receives. The question now is how people will determine what caused the bad outcome.

The thing to focus on here is the general principle that people tend to consider counterfactuals in which events of *unusual* types are replaced by events of *usual* types. Since the reviewer is taking a very unusual approach to the manuscript, people immediately consider the counterfactual in which he takes a more usual approach. That is, they consider a counterfactual of the form:

(1a) If the reviewer had applied a more ordinary standard to the manuscript . . .

The evaluation of this counterfactual then leads (in accordance with whichever theory turns out to be correct) to a judgment that the bad outcome was caused by the reviewer's unusual standard.

But there is also another aspect to the situation. We have been assuming that the reviewer does not reject absolutely all manuscripts and that there was therefore some way of writing the paper that would have led it to be accepted. (For concreteness, we might suppose that the paper would have been accepted if it had offered fulsome praise for the reviewer's own prior work.) Suppose, then, that people began wondering whether the bad outcome was actually caused by something about the paper—namely, the fact that it wasn't one of the .01% of papers that the reviewer would have accepted. To address this question, they would have to consider the counterfactual:

(1b) If the paper had been one of the .01% that fulsomely praised the reviewer's prior work . . .

But there is no principle that picks out this counterfactual as a relevant one. Hence, the counterfactual is classified as irrelevant, people do not give it any further consideration, and the properties of the paper itself don't end up being regarded as causes of the outcome.

A similar approach can be applied to understanding the role of moral considerations. Our example here involved a professor and an administrative assistant who each take a pen from the receptionist's desk. By the second principle laid out above, people should immediately be drawn to counterfactuals that involve changing bad events to good ones. Thus, they should be drawn to consider counterfactuals of the form:

(2a) If Professor Smith had not taken a pen . . .

And they should thereby end up concluding that Professor Smith's decision to take a pen was a cause of the outcome.

But now suppose people begin wondering whether the outcome was also caused by the administrative assistant's behavior of taking a pen. They would have to consider counterfactuals of the form:

(2b) If the administrative assistant had not taken a pen . . .

But there is no principle that would lead people to classify this second counterfactual as relevant. It is therefore classified as irrelevant and blocked from playing further roles in cognition. Ultimately, people do not end up concluding that the administrative assistant's behavior caused the outcome.

What we have here is a rough sketch of an explanation of the effects described above. Clearly, more work will be needed before this explanation can be considered complete, but it should be possible to see, at least in outline, how the various experimental results are to be explained. Above all, it should be clear that the explanation being offered here departs quite radically from the 'person as scientist' theory. If this explanation is on the right track, people's judgments may sometimes mimic the results of a systematic ANOVA, but the basic logic underlying their responses is fundamentally different from anything one might find in a purely statistical analysis. In particular, it seems that the very same process that allows people's judgments to be affected by statistical information also allows them to be affected by moral considerations.

6. Conclusion

The explanation being offered here has a somewhat unusual character, and it may therefore be helpful to say a few additional words about how it is supposed to work and how it contrasts with other explanations that have been offered for the same phenomena.

In thinking about patterns in folk judgments, researchers are often drawn to a mode of thought that might be called *teleological*. That is, researchers are often drawn to the thought that folk judgments must be serving some sort of purpose in people's lives and that we can gain an understanding of why people make these judgments in the way they do by thinking about how they thereby serve that purpose. This mode of thought is especially tempting in cases, like the one under discussion here, in which people's judgments show highly complex patterns. There is an almost overwhelming tendency to suppose that all of this complexity must have arisen because it helps people to accomplish some important purpose.

It seems clear that this sort of thinking is at work in the 'person as scientist' theory. The basic intuition there is that the point of making causal judgments is to achieve a kind of proto-scientific understanding of the world. If the only considerations relevant to that sort of understanding are the statistical ones, then it is assumed that the underlying competence will only take statistical considerations into account. Any use of other sorts of considerations must involve some sort of interference with the proper workings of the mechanism.

On the view presented here, by contrast, it is somewhat difficult to see precisely what purpose the underlying competence might be serving. Hence, a person might ask: 'Why on earth would someone mix together statistical and moral considerations in this complex way? What possible purpose could all of this processing really serve?' If no answer was forthcoming, such a person might conclude that moral considerations must not be playing a role in the competence after all.

My response to this worry is to reject the whole idea that people's underlying competence should be understood as the optimal way of achieving some particular purpose. After all, it is not as though this competence was designed by an engineer who started from scratch and simply tried to create a mechanism that could do the best possible job of generating causal judgments. On the contrary, the competence is best understood as something cobbled together from parts that originally served a different purpose. (Think of the way people sometimes light a fire by using newspaper as kindling. The newspaper is covered in writing—but not because that writing in any way contributes to the function of lighting fires.)

When we consider the matter from this latter standpoint, it is not at all difficult to see why statistical and moral considerations play the role they do. It is not that these considerations came to play a certain role because they could thereby contribute to the purpose of people's causal judgments. Rather, the use of these considerations is simply built into the fundamental mechanisms that subserve people's counterfactual reasoning. Any aspect of human cognition that makes use of counterfactuals will be affected in some way by the structure of these mechanisms. Since causal judgments make use of counterfactuals, and since moral considerations play a role in the mechanisms underlying counterfactual reasoning, moral considerations end up playing a role in causal judgments as well.

Acknowledgements

I am grateful to Jonathan Weinberg and Hunt Stillwell for their suggestions regarding the basic idea at the root of this paper, to Christopher Hitchcock for numerous conversations regarding causal cognition, and to two anonymous referees for suggestions on an earlier version of this manuscript. Finally, I am grateful to Jim Woodward for going beyond the call of duty to provide extremely helpful in-depth comments on all aspects of the present paper.

References

Alicke, M. (1992). Culpable causation. *Journal of Personality and Social Psychology, 63*, 368–378.
Alicke, M. (in press). Blaming badly. *Journal of Cognition and Culture*.
Blackburn, J., & Hakel, M. (2006). An examination of sources of peer-review bias. *Psychological Science*, *17*, 378–382.
Cole, S., Cole, J., & Simon, G. (1981). Chance and consensus in peer review. *Science, 214*, 881–886.
Collins, J., Hall, N., & Paul, L. A. (2001). *Causation and counterfactuals*. Cambridge, MA: MIT Press.
Cushman, F. (2006), Judgments of morality, causation and intention: Assessing the connections. Unpublished manuscript. Cambridge, MA: Harvard University.
Driver, J. (2008a). Attributions of causation and moral responsibility. In W. Sinnott-Armstrong (Ed.), *Moral psychology, Vol. 2. The cognitive science of morality: Intuition and diversity* (pp. 423–439). Cambridge, MA: MIT Press.
Driver, J. (2008b). Kinds of norms and legal causation: Reply to Knobe and Fraser and Deigh. In W. Sinnott-Armstrong (Ed.), *Moral psychology, Vol. 2. The cognitive science of morality: Intuition and diversity* (pp. 459–461). Cambridge, MA: MIT Press.
Försterling, F. (1989). Models of covariation and attribution: How do they relate to the analogy of analysis of variance? *Journal of Personality and Social Psychology, 57*, 615–625.
Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111*, 1–31.
Hilton, D., & Slugoski, B. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, *93*, 75–88.
Hitchcock, C. (2007). Three concepts of causation. *Philosophy Compass, 2*, 508–516.
Hitchcock, C. (2008). Token causation. Unpublished manuscript. Pasadena, CA: California Institute of Technology.
Hitchcock, C., & Knobe, J. (in press). Cause and norm. *Journal of Philosophy*.
Kahneman, D., & Miller, D. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *80*, 136–153.
Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–210). Cambridge: Cambridge University Press.
Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska Symposium on Motivation* (pp. 129–238). Lincoln: University of Nebraska Press.
Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. In W. Sinnott-Armstrong (Ed.), *Moral psychology, Vol. 2. The cognitive science of morality: Intuition and diversity*. Cambridge, MA: MIT Press.
Lewis, D. (2000). Causation as influence. *Journal of Philosophy, 97*, 182–198.
McArthur, L. (1972). The how and what of why: Some determinants and consequences of causal attribution. *Journal of Personality and Social Psychology, 22*, 171–193.
Pearl, J. (2000). *Causality: Models, reasoning and inference*. New York: Cambridge University Press.
Read, D. (1985). Determinants of relative mutability. Unpublished research. Vancouver, BC: University of British Columbia.
Roese, N. (1997). Counterfactual thinking. *Psychological Bulletin, 121*, 133–148.
Solan, L., & Darley, J. (2001). Causation, contribution, and legal liability: An empirical study. *Law and Contemporary Problems*, *64*, 265–298.
Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction and search* (2nd ed.). New York: MIT Press.
Woodward, J. (2003). *Making things happen*. Oxford: Oxford University Press.