Philosophy of Scientific Experimentation 4 (PSX4): Abstracts

11-12 April 2014
Center for Philosophy of Science
University of Pittsburgh

Nina Atanasova (Philosophy, University of Cincinnati)
Validating Animal Models

My purpose in this paper is to show that experimental modeling in neurobiology employs a strategy for calibrating animal models to establish the validity of the knowledge claims about human neurological conditions produced on the basis of laboratory animal experimentation. This way of establishing validity of neurobiological experimental knowledge addresses the critique of the practice of strengthening the reliability of experimental protocols which give rise to only locally valid knowledge claims at the expense of validity raised by Sullivan (2007, 2009).

According to Sullivan, contemporary neurobiology is characterized by a multiplicity of experimental protocols used to study presumably identical phenomena. Different neurobiological laboratories tend to use idiosyncratic experimental protocols and procedures. These different experimental protocols and procedures ultimately produce potentially different laboratory effects which are supposed to represent identical natural world phenomena. In Sullivan's view, the assumption that different laboratory effects correspond to identical natural world phenomena is not justified. This precludes the integration of neurobiological knowledge and further its extrapolation to phenomena outside the laboratory. Further, the way to secure the possibility of extrapolation of neurobiological knowledge is by increasing its validity. According to Sullivan, this entails making laboratory animals and environments more similar to the natural world phenomena which they aim to represent. Because natural world phenomena are complex, increasing the validity of their laboratory representations requires making the laboratory models more complex. This prescription, however, goes against the prescription of reliability to make experimental designs simpler in order to secure reproducibility of laboratory effects. The way to tame the tension between validity and reliability, according to Sullivan, is to increase validity by sacrificing as little of reliability as possible. The alternative conceptualization of validity as convergent validity, I offer here, allows for strengthening both validity and reliability which are of equal importance for good science.

I argue that in the development and validation of animal models as tools for neurobiological experimentation, experimental neurobiologists employ a calibration strategy similar to the strategy discussed by Franklin (1997) and Skipper (2004). Animal models in neurobiology include non-human animal organisms or their parts as components of experimental systems which are used to simulate human neurological conditions and disorders. As such they have representational goals and have to be evaluated for their representational fit to the targeted conditions. That is to say, animal models have to be validated. I argue that calibration is among the major strategies used by experimental neurobiologists for establishing what Campbell and Fiske (1959) call convergent validity. Further, establishing convergent validity, which involves reproducing the same experimental effects through different tests, allows experimenters to use multiple simple models instead of building complex models that would resemble more closely

the targeted conditions in their natural environments and occurrences. This is important because keeping the experimental models simple allows for better control and reproducibility of laboratory effects. In this way, reliability does not have to be compromised in order to achieve validity, which is assumed on Sullivan's account of the relationship between reliability and validity.

Calibration of animal models takes three forms: (1) animal models are tested against multiple known factors to confirm that they reproduce known effects; (2) different animal model designs are tested against each other to check whether they produce converging, or compatible, results; (3) identical animal model designs are tested for reproducibility of effects and convergence of results in different laboratories.

I show that (1) is employed in the establishing of the domestic fowl chick model of the anxiety-depression continuum (Warnick, Huang, Acevedo and Sufka 2009); (2) is at the heart of the standard use of test batteries such as the test battery for studying learning and memory (Sweatt 2010); and (3) is used for the purposes of standardization, e.g. Vorhees (1987) and Wahlsten (2001).

The experiments I study all rely on some sort of converging of results produced on the basis of different experimental arrangements. The variations may include different pharmacological substances whose physiological action is well documented in other species or other models, differences in the administering of behavioral tests or the tests chosen for producing potentially converging results. What is important is that producing compatible and converging results on the basis of multiple experimental arrangements strengthens the likelihood of each line of converging results. Those results then validate one another.

These procedures lead to the establishing of convergent validity. Inherent in this process is the requirement for obtaining converging results from at least two different experimental arrangements (Campbell and Fiske 1959). Therefore, a multiplicity of experimental protocols and procedures is beneficial for establishing convergent validity and even though each experimental model individually only has local validity and captures only limited aspects of the studied phenomena, the integration of converging results produces knowledge which extends further than each individual laboratory context of each individual laboratory model.

Ultimately, I argue that the inherent contradiction between the prescriptions imposed on experimental design by reliability and validity as identified by Sullivan (2007, 2009) dissolves when the validity of an animal model as a representation of a human condition is construed as convergent validity as opposed to external validity. This allows me to respond to Sullivan's (2007, 2009) worry that using multiple simple laboratory models of presumably identical phenomena precludes the validation and integration of neurobiological knowledge claims about human conditions produced on the basis of animal experimentation.

References:

Campbell, D. T. and Fiske, D. W. (1959). Convergent and Discriminant Validity by the Multitrait-Multimethod Matrix. Psychological Bulletin, 59(2), 81-105.

Franklin, A. (1997). "Calibration". Perspectives on Science 5.1: 31-80.

Skipper, R. A. Jr. (2004). "Calibration of Laboratory Models in Population Genetics". Perspectives on Science 12.4: 369-393.
Sullivan, Jacqueline A. (2007). Reliability and Validity of Experiment in the Neurobiology of Learning and Memory. Dissertation.
Sullivan, Jacqueline A. (2009). "The Multiplicity of Experimental Protocols: A Challenge to Reductionist and Non-reductionist Models of the Unity of Neuroscience." Synthese 167, 511-539.
Vorhees, C.V. (1987) "Reliability, Sensitivity and Validity of Behavioral Indices of Neurotoxicity." Neurotoxicology and Teratology 9, 445-464.
Wahlsten, D. (2001). "Standardizing tests of mouse behavior: Reasons, recommendations, and reality". Physiology and Behavior 73, 695-704.
Warnick, J.E., Huang, C. J., Acevedo, E. O., and Sufka, K. J. (2009), "Modeling the anxiety-depression continuum in chicks", Journal of Psychopharmacology 23(2), 143 – 156.

Nora Mills Boyd (History and Philosophy of Science, University of Pittsburgh)
Equivalence Principle Tests

The Equivalence Principle (EP) supposedly plays a central role in characterizing the theory of general relativity (GR) and in particular, the geometric interpretation of gravity. Moreover, null results from decades of experimental research looking for violations of the EP contribute support to the claim that GR has passed all experimental tests so far.

However, there is significant disconnect between formulations of the principle in theoretical physics and philosophical literatures on one hand, and experimental practice on the other. Precise formulations of the principle typically apply strictly to abstract or highly idealized systems involving force-free 'test' bodies, perfectly homogeneous gravitational fields, and infinitesimally small regions. Indeed, one of the tasks adopted by philosophers has been to provide a formulation of the principle from which non-relativistic concepts—such as gravitational and inertial mass, acceleration, and gravitational fields—have been excised.

In striking contrast, EP experimentalists measure and manage forces, account for tidal effects, and capitalize on (or compensate for) the inhomogeneous gravitational environments of real, physically extended, laboratories. Furthermore, these tests are typically described by experimentalists in Newtonian terms. So what do null results from experiments searching for EP violations actually tell us about GR?

The present project aims to bridge the gap between the conceptual foundations of GR and the experiments that supposedly support that theory. To do this, I emphasize the importance of measuring and engineering particular gravitational environments in EP experiments.

I consider recent Lunar Laser Ranging (LLR) and torsion pendulum experiments, which together have been used to test the Strong Equivalence Principle (SEP). In the Apache Point Observatory Lunar Laser-ranging Operation (APOLLO), a SEP violating signal would be observed as a periodic deviation in the predicted lunar orbit. Such a signal would be expected if gravitational interaction varied according to gravitational self-energy. Since the gravitational self-energy of the Earth and the Moon differ, such a violation would manifest as a difference between the

Earth's gravitational acceleration with respect to the Sun compared to the Moon's.

However, the two bodies also differ significantly in composition—the Earth has a large iron-nickel core. In order to rule out the possibility that a composition-dependent EP violating effect *cancels* one due to differing gravitational self-energies (producing a net null result in the LLR experiment), the Eöt-Wash group looked for EP violations using masses of similar composition to the Earth and Moon, but with relatively insignificant gravitational self-energies in a laboratory torsion pendulum experiment. In that context, a signal would be an excess torque generated by the different accelerations of the masses, which would be detected as an extra change in the orientation of the pendulum.

Very sophisticated control of uncertainties is required in both experiments. APOLLO aims at measuring the distance between the centers-of-mass of the Earth and the Moon to *1 mm* precision by measuring the time-of-flight of laser pulses generated on the surface of the Earth and returned by retroreflectors located on the surface of the Moon. Among other techniques, this experiment requires modeling or measuring the effects of a host of dynamical processes and subtracting these from the observational data. For example, solid-Earth tides, atmospheric diffraction, gravitational attraction of planets and asteroids in the solar system, torque due to the rotation of the fluid lunar core, and thermal expansion of the retroreflectors are all be modeled in order to make this precision measurement. The gravitational gradients at the observatory site are also affected by minor changes in the local environment, such as moving the telescope dome, which have a subtle effect on the lunar range data.

In fact, the most important source of uncertainties faced by the Eöt-Wash group is the inhomogeneous gravitational field present in their laboratory. One task of these experimenters is to both understand, and in some cases to physically compensate for, their complex gravitational environment. The Eöt-Wash group has measured location-specific gravity gradients, the effects of which they then canceled with "compensators", that is, large masses strategically placed around the measuring apparatus in the laboratory. Over the years, this group has had to learn to account for variations in not only the ground water retention of the nearby hillside, but also the surrounding flow of car traffic and human activity.

These techniques demonstrate that experimentalists understand the EP as meaningfully applying to real physical systems rather than only idealizations or mathematical objects. In fact, these experiments have been carried out using extended masses subject to complex forces, in inhomogeneous gravitational environments. So perhaps it is not surprising that in the experimental physics literature, the EP is often stated imprecisely in Newtonian terms and bears little to no resemblance to formulations found in the theoretical or philosophical literature. But as a result, the connection between the results of such tests and GR, especially the geometric interpretation of gravity, is usually not transparent. Nevertheless, the experimentalist authors often *do* mention the conceptual content of GR when explaining the motivation for conducting their research, indicating the need for such a connection to be made.

I suggest that the results of EP experiments thus far are best understood as supporting the following: once tidal effects, inhomogeneous gravitational gradients and other expected effects such as those due to electromagnetism are either subtracted from observational data or physically

suppressed, there are no *further* differences between the accelerations of masses with different compositions, velocities, gravitational self-energies, in a given gravitational environment. These results are predicted by the geometrical interpretation of gravity and have excluded numerous proposals for new physical interactions.

In other words, the EP can (and should) be understood as applying to physical systems of the sort employed in EP experiments like APOLLO and the Eöt-Wash torsion balance tests. The results of such tests do support the geometric interpretation of gravity. However, this connection would be greatly clarified in both the conceptual foundations of GR literature as well as that of the EP experimentalists if the the relation between theory and experiment were made more explicit.

Carl Craver (Philosophy, Washington University in St. Louis)
Thinking about Interventions: Optogenetics, Experiments, and Maker's Knowledge

The biological sciences, like other mechanistic sciences, comprise both a modeler's and a maker's tradition. The aim of the modeler in biology, in the narrow sense intended here, is to describe correctly the causal structures, the mechanisms, that produce, underlie, maintain, or modulate a given phenomenon or effect seen in the living world. Such models are expected to save the phenomenon tolerably well (that is, to make accurate predictions about it) and, in many cases, to correctly represent the components and causal relationships composing the mechanism for that phenomenon. The aim of a maker, in contrast, is to build machines that produce, underlie, maintain, or modulate effects we desire. Such maker's knowledge might be deployed in the service of modeler's knowledge, as when engineering triumphs become the next generation of experimental intervention and detection, or it might be deployed for good or ill to serve our needs.

The works of maker and modeler alike depend fundamentally on the ability to intervene into a system and make it work contrary to how it would work were it left to its own devices. The aim of this essay is to identify some dimensions progress (or at least difference) among different means of intervening into biological systems for these modeling and making objectives.

I use the recent development of optogenetics as an example to illustrate these diverse dimensions of progress and difference in intervention techniques. Optogenetics is a kind of genetic manipulation that makes neurons responsive to light. Karl Deisseroth and colleagues published the first paper using optogenetics in 2005. In 2010, Nature Methods named optogenetics the Method of the Year. Science that year included it among the Top Breakthroughs of the Decade. At the time of writing, hundreds of papers using optogenetic interventions have been published in the highest profile journals in neuroscience. It is widely acknowledged, in other words, that optogenetics constitutes an advance in our ability to intervene into the brain. By looking at how researchers justify this new intervention technique, we gain some insight into the requirements that researchers place on interventions, the arguments by which intervention techniques are validated, and the dimensions along which one intervention technique might be said to improve upon another.

Optogenetics allows researchers to control electrophysiological properties of neurons with light. Researchers insert bacterial genes for light-sensitive ion channels into target cells in a given

brain region. They couple these genes to gene regulatory units that ensure the gene is expressed only in specific cell types. The virus by which this genetic construct is inserted into cells commandeers the cell's protein synthesis and delivery mechanisms to assemble the channels and insert them into the cell's membrane. The researcher then inserts a fiber optic cable into the brain near the region of interest. Light delivered through the cable activates the newly inserted channels. The channels open, allowing ions to flow across the membrane. This ionic current can be used to raise or lower the neuron's membrane potential, and so to modulate or produce electrophysiological signals.

To bring out the advantages of this new intervention technique, I first present a standard schema for thinking about causal experiments. Then I discuss twelve dimensions of progress or difference in the ability to intervene into brain function. For many of these dimensions, what counts as progress can be specified only within the context of a given experimental or practical objective. Nonetheless, by exploring some of the qualities that distinguish intervention techniques from one another, we get a feel for the epistemological principles that guide the assessment of progress in intervention. To catalogue such dimensions does not itself amount to an epistemology of intervention. For such an epistemology, this largely descriptive approach should be supplemented with a normative framework showing how these twelve dimensions of virtue make a difference to causal inference. Here I take some preliminary steps in that direction, but my primary objective is to simply frame some of the most salient dimensions of evaluation in a way that makes transparent where such justificatory arguments might be focused. I close by reflecting on some points of overlap and difference between the ways that makers and modelers think about the epistemology of intervention.

Kathleen Creel (Philosophy, Simon Fraser University)
Machine Learning as Experiment

Although new experimental tools are commonly developed in the course of scientific inquiry, entirely new types of experimental techniques are rare. In this paper, I will suggest that machine learning, a branch of artificial intelligence research that focuses on algorithms which can improve their performance at tasks over time, and its algorithms are not mere detectors of patterns in data, as would be implied by recent papers by McAllister on the meaninglessness of said patterns. Using the work of Bogen and Woodward on data and phenomena, I will argue instead that techniques in machine learning such as genetic programming allow us to use patterns in data to get at genuine phenomena, and in a way that allows for discovery of a broader range of phenomena than traditional techniques of investigation.

Given the increasingly common use of large datasets in scientific inquiry, such as the massive amounts of data produced by the Large Hadron Collider or used by government agencies for use in social science, the relationship between the datasets and the phenomena and between the computational techniques used to find patterns in the phenomena have not been sufficiently analyzed by philosophers. The recent work on computer modeling provides a valuable analogy, but the computational analysis of large datasets is not the same as computational modeling of weather systems, movements of tectonic plates, or proposed buildings.

The challenges and problems posed by large data sets, especially ones of pattern recognition and

optimization, are often best solved by writing a computer program that can "learn", training it, and setting it loose on the data set. This set of techniques, from the field known as machine learning, can produce programs that perform much better than a static program written by human hands, one whose performance does not improve with experience. Some types of machine learning algorithms allow the text of the program itself to change over repeated iterations, given defined goals. One such way to help a program to change is to allow similar programs to re-combine and produce new programs, as if through sexual reproduction. As with sexual reproduction, each new program will contain elements of both of its "parents" and also an element of randomness. This is the machine learning technique known as genetic programming.

The ability of genetic programming to change over time can free the researcher from the strictures of habit and prior bias, and it can also help deal with data about which we are massively uncertain. This uncertainty can arise in at least two ways. First, there are uncertain or underspecified domains, problems about which we know so comparatively little about how to solve the problem that it is better to let the learning algorithms start afresh than to pollute the process with guesses. Second, there are domains with data so massive that that a human researcher would have no hope of holding all the relevant information in mind.

In these cases, a two-stage machine learning process may be valuable. The first pass of a machine learning problem may be to identify the possible variables at hand, based either on a test data set or on a criterion of evaluation for what would count as a variable. Then once variables are selected, a second set of runs might determine the best values for those variables. Machine learning techniques allow us to be initially neutral as to details of the model, so that a better fit between model and phenomena can be achieved.

I will suggest that machine learning can be a way of getting at phenomena, not just picking out patterns in data; that it is a form of experimentation; and that genetic programming specifically can go beyond experimentation to allow us to discover phenomena for which we had not known to look.

Machine learning and other computational techniques can help us explain and experiment upon our phenomena, but they can also help us discover new phenomena. Machine learning can do this in two ways. First, the learning aspect of the process can develop new ways of finding existing types of phenomena that may be sufficient improvements to help us find new phenomena. Furthermore, in conducting their experiments, researchers can be tied to certain ways of doing things, such as mathematical techniques or established procedures. Tools such as genetic programming allow the algorithm itself to change and grow, often ending up with new and improved algorithms.

Machine learning can help us discover new phenomena for which we did not explicitly know to look. Allowing machine learning algorithms to search a dataset for data that matches our end goal can be a very broad field indeed. If the algorithm first chooses its own variables and then proceeds to optimize its algorithm based on those variables, the researcher's control is only in writing a disciplined program that will learn well and in setting the goal function by which the success of each stage will be measured. The actual phenomena identified may not be the ones

that the researchers had in mind. Machine learning can therefore be used to pick out phenomena in the discovery phase, as well as to experiment and to explain.

Peter Distelzweig (Philosophy, Western Michigan University)
William Harvey's Really Good (Aristotelian, Socratic, Whewellian) Experiments

William Harvey (1578-1657) has long been hailed as an important early 17th century proponent of experimental methods. Indeed, even in his own lifetime, his explicit emphasis on and successful use of (interventionist) vivisection were noted and lauded by many. His short, cogent, carefully crafted articulation of the experimental justification for his radically new theory of the motion of the heart and the blood in the *De motu cordis* (1628) convinced many (though not all). It led Thomas Hobbes to place him alongside Copernicus and Galileo as a founder of genuine natural science and to note that, to his knowledge, Harvey was the only person that was able to establish a new doctrine in his own lifetime.

In this paper, I articulate four interconnected dimensions along which Harvey's experiments are Really Good Experiments. They are evidentially potent, theoretically fecund, technically expert, and methodologically sophisticated. The theoretical fecundity of Harvey's experiments is due to his imbedding them within (a) a careful articulation of the difficulties and incoherencies of the received, Galenic theory of the physiology of the cardiovascular system, (b) concrete, simple models of his proposed alternative, (c) and a shared (largely Galenic) conceptual framework in common to both theories. (In this way, Kuhnian problems of incommensurability and diverging meta-theoretical values are to a large extent defused.) The technical excellence of his experiments lies not so much in the establishment of precise measurements, as in the personal, manual skill in dissection involved in isolating significant results. The evidential potency of his experiments can be understood in terms of Whewellian induction, and even consilience of inductions. Harvey shows that a diverse group phenomena—both familiar ones and new, experimentally produced ones—can be seen as "one fact," if his theory is accepted. Finally, Harvey's experiments are methodologically sophisticated in (at least two) ways. His experimentation is guided self-consciously by an explicit method and that method represents a creative and effective response to characteristic difficulties in arriving at knowledge of physiological function. Interestingly, this methodology is and is understood by Harvey to be due to Aristotle, and ultimately to (Plato's?) Socrates. Thus, he encapsulates his evidentially motivated comparative method in the phrase, "The Rule of Socrates."

After articulating these dimension of Harvey's Really Good Experiments, I attend especially to the final, methodological dimension, tracing Harvey's understanding of its Aristotelian and Socratic roots, drawing on Harvey's lecture notes from (roughly) the decade leading up to the publication of *De motu cordis* in 1628, and its presence in that seminal text. I argue that, ultimately, it is this method that accounts for the excellence of Harvey's experiments along the other three dimensions. The examination of such an historical case, besides having its own philosophical interest, also sheds light on the origins of privileging experiment in the natural sciences.

Melinda Fagan (Philosophy, Rice University)
Crucial Stem Cell Experiments? An Objection to the Uncertainty Principle for Stem Cells

This paper responds to an objection to my recent work on stem cell experiments (Fagan 2013a, 2013b). The objection is to my thesis that empirical claims about stem cell capacities are inherently uncertain, due to features of the stem cell concept together with general facts about experiments in stem cell biology. Stem cells are defined as cells capable of both self-renewal (reproduction yielding offspring that resemble the parent) and differentiation (change in cell traits to yield more specialized cells). So the stem cell concept is relational, in that its application depends not only on cells' intrinsic properties but also on relations between parent and offspring cells.

On my view, the general stem cell concept is treated as an abstract model with variables corresponding to organismal source, cell lineage, cell traits and a temporal duration of interest. Experimental methods for identifying stem cells specify values for these variables. Substantive claims about stem cells, therefore, must be understood in terms of experimental methods used to identify the stem cells at issue. But these methods are subject to an evidential constraint: self-renewal and differentiation potential cannot be experimentally measured for a single cell. To determine a cell's differentiation potential, that cell is placed in an environment conducive to differentiation, and its descendants measured. To determine its self-renewal ability, the cell is placed in an environment that inhibits differentiation, and its descendants measured. It is not possible to perform both experiments on a single cell. So the two defining stem cell capacities cannot be measured for a single cell and, therefore, stem cells cannot be identified at the single-cell level.

Laplane (ms.), as well as several scientists in conversation, have objected that some stem cell experiments do measure self-renewal and differentiation potential at the single-cell level. These experiments use whole animals (inbred mice) to measure stem cell capacities; i.e., in vivo rather than in vitro methods. The method, briefly, is to transplant a single hematopoietic (blood-making) stem cell into an inbred mouse whose immune system was previously removed by radiation (e.g., Spangrude et al 1988, Kondo 2010, Naik et al 2013). Any immune cells that subsequently appear in that animal must therefore be derived from the transplanted cell. Because immune cells have very short lifespans, self-renewal is required to maintain an organism's immune system over time. It follows (so the objection goes) that both self-renewal and differentiation potential can be measured by in vivo experiments of this kind. Such single-cell transplantation experiments are a "gold standard" for one branch of stem cell biology (Melton and Cowan 2009). So the objection is a significant one. This paper offers a response.

I show that in vivo single-cell transplantation experiments do not overcome the evidential constraints discussed above. Laplane's objection is blocked at three points. First, the experiments do not demonstrate that the transplanted cell itself is capable of self-renewal, but only that self-renewal occurs somewhere in the lineage originating with the transplanted cell. It is possible, for instance, that the transplanted cell immediately divides to produce more specialized progenitors capable of self-renewal, which in turn give rise to distinct blood cell lineages. For in vivo stem cell experiments, unlike in vitro, measurement of self-renewal is indirect, inferred at the final stage of testing rather imposed at the outset by experimental design. Second, single-cell

transplants do not unequivocally establish the transplanted cell's differentiation potential. The experiment shows that the transplanted cell can give rise to immune cells in the context of a particular (extensively manipulated) animal. But this does not tell us anything about its potential in other contexts. Third, single-cell transplants are usually inferred from limiting dilution assays, which require a population of candidate stem cells assumed to be identical. This 'homogeneity' assumption is relative to experimental context; specifically the cell characters measured. But we cannot know in advance which cell characters are the right ones for identifying the kind of stem cell in question.

This final point involves an important clarification of my view. I do not claim that experiments involving self-renewal and differentiation cannot be performed on a single cell (which is clearly contradicted by single-cell transplantation experiments). My claim is rather that experiments aimed at identifying stem cells and their capacities cannot unequivocally demonstrate those capacities as defined by the prevailing stem cell concept. This thesis concerns only those experiments that aim to tell us what cells qualify as stem cells (under particular experimental conditions). If we can assume at the outset that a given cell is a stem cell (as identification of the transplanted cells as HSC implicitly does), the evidential constraint does not apply. However, this assumption is unjustified in stem cell research today, because we do not have a way independent of these very experiments to identify stem cell properties and capacities. Characterizing transplanted single cells as blood-making stem cells (HSC) is putting the evidential cart before the horse, so to speak. The tendency to hypothetically characterize cells in this way appropriate in theoretically-driven, but not experimentally-driven, scientific fields.

The overall lesson here is not full-blown skepticism about stem cell capacities, but methodological caution. One of my conclusions is that we can use "single-cell standards" to get good evidence about stem cells in particular experimental contexts (Fagan 2013a, 2013b). Single-cell in vivo transplants are one such standard. So I am happy to acknowledge the value of these experiments for stem cell research. But they do not escape the basic evidential constraint of the 'uncertainty principle' for stem cells.

References:

Fagan, MB (2013a) Philosophy of Stem Cell Biology. London: Palgrave Macmillan.

Fagan, MB (2013b) 'The stem cell uncertainty principle' Philosophy of Science 80: 945-957.

Kondo, M (ed.) (2010) Hematopoietic Stem Cell Biology. New York: Humana Press.

Landecker, H (2007) Culturing Life. Cambridge: Harvard University Press.

Laplane, L (ms.) Cellule souche cancéreuses: ontologies et therapies. Ph.D. Dissertation, Université Paris Ouest Nanterre La Défense and Sorbonne Université. To be published (in translation) by Harvard University Press.

Melton, DA, and Cowan, C (2009) 'Stemness: definitions, criteria, and standards.' In Lanza, et al (eds.) Essentials of Stem Biology, 2nd edition. San Diego, CA: Academic Press, pp. xxii-xxix.

Naik, SH et al (2013) 'Diverse and heritable lineage imprinting of early haematopoietic progenitors.' Nature 496: 229-232

Spangrude, G., Heimfeld, S., and Weissman, I.L. (1988) 'Purification and characterization of mouse hematopoietic stem cells.' Science 241: 58-62.

Paula Grabowski (Biological Sciences, University of Pittsburgh)
Perspectives on RNA and the Evolution of Biological Catalysis and Proteomic Diversity

This talk will highlight experiments that led RNA biologists to escape the mindset of the Central Dogma hypothesis and open the door to the new frontier of catalytic RNA. The accidental discovery of the enzymatic functions of the Group I intron of *Tetrahymena thermophila* prompted us to expect the unexpected about the roles of RNA within the essential workhorse machineries of protein synthesis and RNA splicing. The finding that contemporary RNA molecules can have dual functions encompassing genetic storage and catalysis has inspired ideas for plausible pathways operating in an evolutionary time frame that may explain the origin of RNA-based viruses, mobile introns, and RNA-protein machineries. The RNA World hypothesis paints this picture in broad strokes making predictions that are experimentally accessible. The big surprise is that single strands of RNA can fold into remarkably intricate secondary and tertiary structures that provide active sites for substrate binding, or interaction sites for protein assembly. The close partnership between RNA and protein cofactors is of broad importance in the phenomenon of alternative RNA splicing, which generates the vast diversity of proteins in human cells.

Spencer Hey (Biomedical Ethics, McGill University)
Uncertainty, Underdetermination, and the Units of Clinical Translation

What makes a good clinical biomarker experiment? The promise of personalized medicine, which depends on the development of high-quality biomarker diagnostics, hinges on the answer to this methodological question. Indeed, the goal of personalized medicine is to equip the health-care system with an array of clinically validated diagnostics, each of which would allow physicians to test their patients for the presence or absence of a particular biomarker, and then use these results to guide decision-making about the appropriate course of treatment for that particular patient. If successfully implemented, these biomarker diagnostics would potentially save the health-care system billions of dollars and prevent needless patient suffering due to futile interventions.

Unfortunately, as Hayes et al. (2013) and numerous commentators in the journal "Clinical Trials" (Oct. 2013) have recently emphasized, the quality of most biomarker studies is quite low. This has lead to a vicious cycle wherein evidence from biomarker studies is poorly valued, biomarker research is poorly funded, the costs of diagnostics are not reimbursed by health-care providers, and this leaves little incentive to improve the quality of evidence. Although these commentators have discussed some of the technical and social factors that contribute to the problems with biomarker experiments, the more fundamental philosophical issues remain unexplored.

In particular, biomarker experiments challenge the standard model of clinical translation---that is, the process of developing new therapeutic interventions from the laboratory bench to the clinical bedside. The standard model assumes that the relevant unit to be translated is an individual agent, such as a molecule or drug. However, this model does not accurately characterize biomarker development. Consider, the case of the anticancer therapy, temozolomide, which is approved for use in patients with malignant glioma whose tumors test

positive for the methylated-MGMT biomarker: The question that needs to be answered in these biomarker experiments is not simply whether temozolomide is effective, since we already know that it can work. Rather, the question is: What are the necessary and sufficient conditions for leveraging a methylated-MGMT diagnostic in order to maximize the therapeutic benefit of temozolomide in glioma patients? The relevant unit of translation is therefore not the drug per se, but a "therapeutic ensemble," which includes a sensitive and specific assay, a rigorously defined patient population, a particular drug dose and schedule, various co-interventions, delivery techniques, and so on. A successful biomarker translation depends upon investigators discovering the correct values for each of these parameters.

In this paper, I begin by showing how this shift in units from individual agents to therapeutic ensembles further complicates the problem of underdetermination. In the traditional model of medical research and drug development, there is a single hypothesis about the effectiveness of the experimental treatment and a single theory of disease mechanism that drives the research program. A new drug that is successfully tested and implemented in the clinic is taken to confirm both the hypothesis of its effectiveness and the underlying theory. Whereas a drug's failure is attributed to either a problem with the theory, a faulty auxiliary hypothesis, or an operational error in one or more of the experiments.

Biomarker testing, in contrast, has at least four other dimensions of uncertainty: (1) the mechanistic theory explaining the biomarker's relationship to the drug and disease; (2) the predictive capacity of the biomarker to identify the clinically relevant population (i.e., "clinical validity"); (3) the mechanistic theory of the diagnostic assay (or assays) used to identify the biomarker; and (4) the accuracy with which the assays classify patients as either biomarker-positive or -negative (i.e., "assay validity"). These additional uncertainties not only mediate the interpretation of the study results, but also judgments of study quality. For example, biomarker studies rarely report quantitative misclassification rates for the assay diagnostics used to determine the patient's (or more accurately, the tissue sample's) biomarker status. And yet, without this information, we cannot know whether the assay used in a study has adequately demarcated the biomarker positive and biomarker negative populations, rendering suspect any conclusions about the biomarker's clinical validity and utility. Similarly, for biomarker studies that use multiple assays, any disagreement in classification between the assays calls into question the posited theoretical relationship between the assay, the biomarker, and response to the drug.

Ultimately, I argue that this more complicated epistemology has important consequences for how we should understand what makes a good clinical biomarker experiment: (1) It amplifies the need for robustness analyses across experiments in order to ensure that ensemble parameters are discovered efficiently; and (2) it demands that there is more strategic coordination among research actors in order to address lingering parameter uncertainties and prevent duplicate or unnecessarily risky investigations.

Andréa Loettgers (Philosophy, University of Geneva)
Modeling/Experimenting? The Synthetic Strategy in the Circadian Clock Research

The similarities and differences of modeling and experimentation have become a subject of intensive discussion. The standpoints taken in this discussion are studied through the case of combinational modeling in synthetic biology. In combinational modeling the experiments on model organisms and mathematical models are triangulated with a new type of model—a synthetic model. This strategy is due to the characteristic constraints of these three epistemic activities that are, in turn, related to their different materialities. Synthetic modeling shows that the question of materiality should not be reduced to the "the same stuff". The mechanism of interaction is also crucial.

Irina Meketa (Philosophy, Boston University)
How Parsimony Biases Experimental Design in Comparative Cognition

In this paper, I illustrate how an undefended preference for parsimony imports theoretical biases into the development of experimental research programs and cognitive models (models of what processes and mechanisms may be responsible for observed behavior) in comparative cognition research. Parsimony is widely considered to be a virtue of scientific theories, experiments, and models. Less widely appreciated, however, is the fact that a preference for parsimony can bias experimental investigation. In the case of comparative cognition, this bias results in a tendency to under-attribute putatively sophisticated cognitive abilities to nonhuman animals – a claim that I defend in detail elsewhere (Meketa 2014).

My discussion proceeds as follows. I first provide a brief historical overview of the motivations underlying the preference for parsimony in order to illustrate how this value has shaped the theoretical assumptions that guide comparative cognition. I focus primarily on two competing types of accounts: 'associative' accounts (which rely on learning through repeated exposure to paired stimuli) and 'metacognitive' accounts (which posit a representation of one's own or others' mental states). Associative mechanisms are presumed to be simpler, or more parsimonious, than metacognitive explanations, though plausible defenses for this presumption are lacking. What matters in the present context, however, is not the justification of the parsimony-based preference for association over metacognition, but rather the effects that this preference has on the development of experiments in comparative cognition. Next, I show that one crucial consequence is that the preference for parsimony has bestowed unwarranted evidentiary weight to putatively simpler cognitive models in such a way as to affect the developmental trajectory of one set of behavioral experiments, which I examine in this paper. These experiments test for the presence of metacognition in rats, and were conducted by Jonathon Crystal and Allison Foote over the course of a number of years.

Foote and Crystal (2007) began with an experiment from which they concluded that rats are capable of uncertainty-monitoring – a form of metacognition. They tested rats in a duration-discrimination test, where the animals were tasked with categorizing short, long, and ambiguous tones as either "short" and "long." When presented with a third option – to decline a test – the rats consistently opted to decline the ambiguous ("difficult") tests but not the unambiguous tests. Crystal and Foote concluded that their rats' behavior demonstrated knowledge of uncertainty – a

sign of metacognition.

However, Crystal and Foote (2009) withdrew their support for their earlier conclusion once a novel cognitive model became available. On this model, Smith et al. (2008) offer a putatively simpler "response-strength model" as an associative alternative to the metacognitive explanation of the rats' performance on the duration-discrimination tests. Because Crystal and Foote (2009) agreed that association is simpler than metacognition, they conceded that the mere availability of an associative model capable of simulating their rats' behaviors defeats their metacognitive hypothesis. In a recent paper, they write: "Clearly, putative evidence for metacognition in rats is critically undermined when a non-metacognition model can produce the observed pattern of behavior" (Foote and Crystal 2012, 188). In response, they have attempted to structure future physical experiments so as to rule out the Smith et al. (2008) explanation, though without success.

The consequence for the direction of research is stark: When competing explanations – the metacognitive explanation and any "simpler" explanation – are underdetermined by the behavioral evidence, the metacognitive explanation has the burden of proof. Since associative explanations are considered incompatible with metacognitive explanations, any associative model that makes the same predictions as the metacognitive model will be preferred without further evidence that the mechanism it postulates is actually responsible for the behavior. The parsimony-based commitment to associative hypotheses over metacognitive hypotheses determines the course of the research programs on metacognition. According to Foote and Crystal, the introduction of putatively simpler models "necessitates the development of new, innovative methods for metacognition" (Crystal and Foote 2009, 1). However, there is no evidence that the simplest of two phenomenological models should be the most likely to be true.

Now suppose that metacognition were the default hypothesis. Then, the burden of proof would require cognitive modelers such as Smith et al. (2008) to produce a model that made more correct predictions than the metacognitive model. In order to test these predictions, researchers who wished to defeat the metacognitive explanation would need to devise experimental protocols that differentiated between a specific associative model and the metacognitive model. Instead, behavioral experiments must, according to the comparative psychological orthodoxy, continue to further refine experimental protocols for metacognition until no associative explanation is available. This example illustrates how a theoretical value can surreptitiously influence the direction of experimental research programs by granting potentially illicit evidentiary status to simple cognitive models and computer simulations.

The Foote and Crystal example is not unique: cognitive modeling and computer simulations of animal behavior are becoming increasingly popular in comparative cognition. For example, van der Vaart et al. (2012) offer a computer simulation of food-re-caching by Western scrubjays that is based on just one rule: "re-cache food more when more stressed." This simulation explains re-caching behavior in allegedly simpler terms than the alternative metacognitive explanations, which suggest that scrubjays are aware of threats to their food caches. Similarly, Bell and Pellis (2011) were able to simulate the theft-aversion behaviors among rats using a single rule: "keep the distance between your nose and the nose of the other rats constant." Both van der Vaart et al. and Bell and Pellis were able to simulate the behaviors and hormonal stress levels of real animals

using computer simulations. While they were each cautious in drawing implications from their successes, both sets of researchers suggested that the simplicity of their models provided evidentiary support to their hypotheses. As long as the simplicity of these models continues to count as an epistemic virtue, more scientists will seek simpler models and more researchers will need to modify their physical experiments in order to respond to the perceived challenges posed by these models.

In the end, my analysis shows how unexamined theoretical values can covertly shape the evolution of experimental research by giving epistemic weight to cognitive models and computer simulations on the sole basis of their conformity with the theoretical value. More broadly, the Foote and Crystal case illustrates how a closer scrutiny of the role of epistemic values in cognitive modeling and computer simulations can shed light on the relationship between experimentation and theory.

References:

Bell HC, Pellis SM (2011) A cybernetic perspective on food protection in rats: simple rules can generate complex and adaptive behavior *Animal Behaviour* 82: 659-666.

Crystal JD, Foote AL (2009) Metacognition in animals. *Comparative Cognition & Behavior* 4: 1 – 16.

Foote AL, Crystal JD (2007) Metacognition in the rat. *Current Biology* 17: 551 – 555.

Foote AL, Crystal JD (2012) Play it again: a new method for testing metacognition in animals. *Animal Cognition*. 15:187–199

Meketa I (2014) A critique of the principle of cognitive simplicity in comparative cognition. *Biology and Philosophy* doi: 10.1007/s10539-014-9429-z

Smith JD, Beran MJ, Couchman JJ, Coutinho MVC (2008) The comparative study of metacognition:sharper paradigms, safer inferences. *Psychology Bulletin Review* 15:679–691

van der Vaart E, Verbrugge R, Hemelrijk CK (2012) Corvid re-caching without 'theory of mind': a model. *Public Library of Science Online* 7(3):e32904. doi:10.1371/journal.pone.0032904

Sandra Mitchell (History and Philosophy of Science, University of Pittsburgh)
On Relations Between Experimental and Representational Models

Giere in 2010 describes an intentional conception of representation in science. That is, models are characterized in part by their intended use. By adding this pragmatic component to an account of models, Giere holds that he can avoid a fictionalist interpretation of the relationship of models to the world.

His positive account is that there is an indirect, imperfect relationship between models and the world, but a connection nonetheless. The connection of principled models (like Newton's laws) is via their test by models of data that are developed from experiment and observation. However, principles, like F=ma, are abstract and to know "where in the world to look to see whether or not the laws apply" (Giere 2004:745) requires introducing specific conditions (one might say, interpretations) that yield a model that is a step towards being tested by empirical observation. These, which Giere calls "representational models" are still abstract, e.g. F=-kx is a specification

of Newton's 2nd law for simple harmonic oscillators, where x is displacement from equilibrium. To be tested, actual springs and masses need to be observed, and the results of those observations (a model developed from the experimental data) can then "test" the principles by means of the intermediate "representational model".

I will argue that that Giere's hierarchical framework for multiple models used in explaining a specific phenomenon does not exhaustively characterize the relationships between experiment, experimental models and representational models. His view is that representational models come from principles via specification, and that experimental models come from observations via abstraction. They meet in the middle, so to speak, to compare the two models and thereby "test" the principles by observations, by comparing the two intermediate models. This he refers to as a hierarchical picture of scientific modeling. Appealing to examples from ab initio and experimental models of protein folding, I will argue that an important relationship between models of data and representational models is not, in fact, hierarchical, but integrative. The practices involved in generating models of protein folding follow Giere's account only in part. Ab initio or all-atom models are specifications of Newtonian Principles. But in constructing predictive models of protein folds the representational models invoke experimental models in a constructive, not confirmational way. Models of protein structure inferred from x-ray crystallographic and nuclear magnetic resonance experiments (experimental models) are used to modify the representational models that then lead to hypotheses/predictions of specific protein structures. This role of experimental models does not conform to Giere's hierarchy.

References:
Giere, R. N. 2004 "How Models are Used to Represent Reality" Philosophy of Science: 742-752.
Giere, R. N. 2010 "An agent-based conception of models and scientific representation" Synthese: 269-281.

Margaret Morrison, Philosophy, University of Toronto
Bridging the Great Divide: Simulation, Experiments, and Validation Experiments

In debates concerning the merits of experiment vs. simulation a sharp distinction is usually drawn giving the former greater epistemic legitimacy than the latter. The basis for the distinction is often grounded in the 'materiality' of experimental investigation as opposed to the formal, abstract nature of simulation. In this talk I want to examine the role of simulation in the Higgs searches at the LHC. What the experiments (ATLAS and CMS) reveal is a reliance on simulation that significantly undermines the type of sharp division between the two that has characterised philosophical debates. Moreover, assessing the legitimacy of simulation involves much more than formal verification of mathematical algorithms; extensive validation experiments are required in order to ensure the accuracy of simulation as part of the overall experimental context. What this interplay between experiment and simulation in the LHC case shows is the necessity of simulation for the discovery of the Higgs boson. In other words, there's no experiment result without simulation.

John Norton (History and Philosophy of Science, University of Pittsburgh)
Is the Replicability of Experiment a Principle of Inductive Logic?

Is the requirement that credible experimental results must be replicable a principle of inductive logic comparable to the principles of deductive logic, such as the law of the excluded middle?

My answer is "no." Mere replication or its failure has no univocal import in science. We can find cases in which successful replication is judged epistemically significant and others in which it is epistemically inert. And we can find cases in which failure of replication is epistemically significant and others in which it is epistemically inert.

No simple principle of replicability can make sense of these cases. Rather, we make sense of the differing import by identifying the pertinent background facts in each case. Identifying these background facts is already sufficient to determine the evidential import of the results in each case. It follows that we do not need an elusive, general principle of replicability, since such a principle is superfluous to the determination of evidential import.

This analysis comes within a larger project of research in inductive inference. Its core idea is that there are no universal principles of inductive inference and no universal schema of inductive logic. That is, inductive logic should not be modeled on deductive logic, which is based on such principles and schema. Rather inductive inferences are warranted by facts that prevail locally.

Paolo Palmieri (History and Philosophy of Science, University of Pittsburgh)
What Makes a Good Experimentalist? Among Other Things, Good Senses…

I explore the idea that a unilaterally mechanistic model of the senses undermines the good experimentalist. Rethinking the achievement of the senses means overcoming the separation of world and thought that has prevailed in Western science.

Emily Parke (Philosophy, University of Pennsylvania)
Experiments, Simulations, and Surprises

There is a general feeling among philosophers of science, and scientists themselves, that experiments have epistemic privilege over simulations. That is, experiments are better and more reliable for generating scientific knowledge and valid inferences about the natural world. This paper focuses on one aspect of that idea: The claim that simulations cannot surprise us the way experiments can. A stronger version of this claim would say that simulations cannot genuinely surprise us at all. More commonly, the claim is that simulations and experiments differ in principle, qualitatively or quantitatively, in their capacity to surprise us (Morgan (2005) and Sniegowski (2013) have argued versions of this claim). I argue that the surprise claim is false as a generalization; there is a limited sense in which there is some truth to it, but regarding only a particular kind of surprise. In any case, surprise is not an in-principle epistemic virtue; its value depends on the context of inquiry.

The intuition behind the surprise claim rests on the following sort of idea: While an experimenter often designs some of her object of study's parts and properties, she never designs all of them,

and in certain cases, like in some field experiments, she designs none of them. In computer simulations, on the other hand, a researcher designs all of the parts and properties of her object of study (a model). This difference is thought to imply that simulations cannot surprise us the way experiments can. There are good historical and Bayesian motivations for regarding surprise as critically valuable for scientific inquiry, and thus for thinking this difference would support the case for experiments' epistemic privilege over simulations.

To show why the surprise claim does not hold as a generalization, I focus on an important difference between two kinds of sources of surprise. The first kind is *unexpected behaviors*: surprising states or phenomena in one's object of study exhibited over the course of studying it. These occur in experiments all the time, but they also occur in simulations all the time. I discuss examples from studies of evolving populations in both wet-lab experimental evolution and agent-based simulations.

It makes sense that experiments and simulations have equal potential in principle to lead to surprises in the form of unexpected behaviors. An experiment starts with an experimental object of study and a protocol; a simulation starts with the object of study (a model) having some initial state and set of transition rules. Both involve observing what happens to the state(s) of that object of study over time. Counter to what people making the surprise claim sometimes imply, a simulationist will not always know everything about her object of study. A straightforward case where she might fail to know everything is when her object of study is a model which someone else wrote. But there are more interesting cases as well, such as models written in high-level programming languages, models whose initial conditions include unintended features or whose transition rules entail unintended consequences, or highly complex models written by teams. In any case, knowing "everything" going into a simulation study about the initial conditions and transition rules is not sufficient for knowing what will happen, just as carefully specifying an experimental system and protocol is not sufficient for knowing the experiment's results. Any study of a system with an initial state and subsequent states has at least the potential to surprise us, because it contains potential sources of unexpected behavior as its states change (or fail to change) over time.

The second kind of source of surprise I discuss is *hidden mechanisms or causal factors.* Unlike unexpected results, hidden mechanisms are sources of surprise which can be said to have "been there all along" in the object of study itself, which a researcher was unaware of when she began studying it. A perfect example is Barbara McClintock's discovery of transposable genetic elements over the course of her study of maize genomes. Hidden mechanisms can be found at different levels of organization in one's object of study; in particular, at (i) the molecular, individual or atomic level, (ii) the level of interactions among individuals or atoms, or (iii) the population or aggregate level. I argue that hidden mechanisms of at least the third sort can be found in simulations as well as experiments, again citing examples from studies of agent-based models. While it seems that experiments give us far more opportunities to uncover hidden mechanisms of the first sort, there are arguably examples where simulations have give us such opportunities as well, in research areas such as physics at the nanoscale.

The upshot of all of this is that it is not true that simulations cannot surprise us the way experiments can, especially not as a generalization across science. Both experiments and simulations have the same potential in principle to give rise to unexpected behaviors, and I give reasons to think that both can also lead to the discovery of hidden mechanisms. Though on this latter point, it still seems right to say that simulations do not contain a specific kind of source of a surprise—namely, molecular-, individual-, or atomic–level hidden mechanisms—as often as experiments do.

I conclude by discussing the implications of this difference. If it is true that experiments can contain at least one kind of source of surprise more often or more consistently than simulations can, this is an important point. But this does not support the idea that we can use the experiment/simulation distinction to make in-principle judgments about epistemic value. Surprise is valuable to scientific inquiry because it is *productive*, in the sense of broadening the scope of inquiry, linking research programs in interesting ways, or opening new channels of inquiry. But the value of productive surprises depends on the context of inquiry, with a key distinction being between strict hypothesis-testing and exploratory research contexts. Surprise plays a key role in exploratory research, but in a strict hypothesis-testing setting, we do not seek surprises; in that context, valid scientific inferences come from showing that we have *eliminated* sources of surprise, in a sense.

References:

Morgan, M. (2005). Experiments versus models: New phenomena, inference and surprise. *Journal of Economic Methodology*, 12(2), 317–329.

Sniegowski, P. (2013). Commentary on the symposium "Simulation vs. Experiment in Evolutionary Biology" at *ISHPSSB 2013,* Montpellier.

Sherri Roush (Philosophy, University of California, Berkeley)
The Epistemic Superiority of Experiment to Simulation

This paper defends the naïve thesis that experiment is epistemically superior to simulation, other things equal, a view that has been resisted by many philosophers writing about simulation. I focus on experiments and computer simulations whose purpose is understanding and predicting phenomena in the actual world. There are three challenges in defending this thesis. One is to say how "other things equal" can be defined, another to identify and explain the source of the epistemic advantage of experiment in a hypothetical comparison so defined. Finally, I must explain why this comparison matters, since it is not the type of situation scientists can expect often to face when they choose experiment or computer simulation (hereafter "simulation").

To define "other things equal" we must say what kind of property counts as other and what is required for those properties to be equal. "Other" in this case is shorthand for "other than those properties that distinguish experiment from simulation as types of method", so to some extent I must take a stand on what distinguishes these methods. I will do that by process of elimination of the things that are evidently similar and must be held equal in my comparison. Since I aim to show that experiment is superior, I will err in the direction of taking simulation to be similar to experiment to the greatest extent possible.

First, in my comparison the two studies must be aiming to answer the same question, say, whether atoms have nuclei. Beyond this, typical experiments and simulations have a great deal in common, as others have discussed. (E.g., Parker 2009, Winsberg 2010) Both methods in the uses I'm focused on employ a stand-in, a study system whose results are to be generalized to a target system. In both cases the justification for that generalization goes by way of establishing relevant similarity between the study and target systems, of whatever sort, by whatever means. Both experiments and simulations are run. That is, they are dynamical processes initiated by the functional equivalent of an ON switch. In both experiment and computer simulation these processes are concrete. In experiment this is obvious; for example, the alpha particles are shot at the gold foil and follow a trajectory dictated by physical law. In computer simulation, the process is a computation governed by dynamical laws encoded in a program. That is, in my view, in perfect analogy to an experiment the computer program constitutes a set of dynamical laws that govern the time evolution of hunks of hardware, typically made of silicon. A program is an abstract entity, but so are the laws of physics. What both sets of laws govern are concrete processes. Both methods are interventions in a broad sense. When the switch is flipped on, an initial state – whether this is flying alpha particles and a sheet of gold of a certain thickness, or numerical inputs and their associated silicon – is set free to do its work according to the laws. Both kinds of studies have outputs at the end of the process that are typically called "data". In both methods the data must be interpreted in order to get results, which in turn are used to justify conclusions about the target system.

In both methods interpretation of the data requires assumptions about the dynamical laws and inputs of the study system. For example, Rutherford and Marsden did calculations to determine precisely how thin a sheet of gold had to be in order for alpha particles to back-deflect if, and only if, the atom has a nucleus. (A thick enough sheet of gold would have sufficient density to yield back-deflection whether atoms have nuclei or not.) What those calculations could contribute to correctly inferring nucleus or not from the outputs of their experiment was only as good as their knowledge of the mechanics of collisions between particles of particular masses and velocities, and their knowledge of the number and masses of protons and neutrons in gold, and of the volume of an atom. Similarly, to interpret the data that comes out of a simulation, scientists must make assumptions about what the program was doing when it manipulated the inputs, which depends on assumptions about those "laws" and inputs. The set of numbers (and associated graph) in the data is a virtual hurricane insofar as the program was successfully computing virtual hurricanes of the sort defined in the set-up.

To hold this interpretational aspect of the two methods equal cannot be to say they are identical: neither numerical inputs nor silicon are identical to alpha particles. But this is not the equality relevant to my overall question. Instead, these aspects must be, and can be, assumed equal by supposing that the scientists are equally justified in their assumptions about the laws and inputs of their respective systems, and so, equally justified in their answers to the interpretive, internal-validity question of what accounts for the data in their respective study systems.

That equality refers to what the experimenter and simulator are justified in believing about what determines their data. What actually determines the data in the two studies and, as a consequence, the extent to which they are justified in believing that what happened in the study system is generalizable to the target system, is the remaining dimension and relevant difference

between the experiment and the other-things-equal simulation. What determines the data in the Rutherford experiment is physical laws governing particle collisions, samples of alphas and gold, and a fluorescent screen. What determines the data in a simulation is the program in the solver and simulacra of alphas and gold samples and a fluorescent screen. The crucial point is this: simulacrum gold cannot yield data even putatively revealing whether the simulacrum gold has a nucleus unless the simulacrum gold is programmed to have (a correlate of) internal structure. Otherwise, the program will give no determination at all of the trajectories of the simulacra alphas once they reach the simulacrum gold.

The internal structure programmed in has to be more than what Rutherford and Marsden assumed about the atom in order to calculate their two possible outcomes, because what they knew did not determine the outcome. Otherwise, they wouldn't have needed to do an experiment. A simulation will have to program in something that does determine whether the (virtual) alpha particles back-deflect, and that will require either begging the question of the study or relying on the results of some equivalent of the Rutherford experiment on gold. The experiment does not need to make those further assumptions that determine the outcome because the alphas and gold do that. Thus, a simulation of equal epistemic force to Rutherford and Marsden's scattering experiment on our chosen question cannot be done, because to give data at all on the question, the simulation would have to model – that is, assume rather than discover – something that is sufficient to determine whether the gold atom has a nucleus. The point is not per se that an experiment's similarity to its target system is material – a dimension of comparison that has been frequently discussed – but rather that if we hold the question constant, a simulation is always strictly one step further removed from the target system than the other-things-equal experiment is in how many layers of assumptions must be made in order to have relevant data be produced at all.

Supposing I am right that experiment is epistemically superior other things equal, why does it matter when our choices in practice are not typically between otherwise equal studies? Obviously, we cannot do a total climate experiment that will tell us what we want to know in time for it to be helpful, or explode nuclear missiles when we have signed a test ban treaty, or deliberately infect human beings with a disease. However, my superiority claim makes a difference in every case to our assessment of what we get out of simulations, and this matters in how we reason about whether to do an experiment or simulation in any case where both are or will reasonably soon be possible. Simulations are typically cheaper than experiments, but my thesis implies that that is never the only consideration. The expense of a given experiment must be weighed against the extent to which it is epistemically superior to a proposed simulation about the same question. Though it cannot be quantified precisely and in cases like those requiring deliberate infection of human beings is completely cancelled by the cost, the mere fact of being an experiment rather than a simulation is always an epistemic advantage. This is why even if WHO and AAAS studies had not detailed specific, beneficial research that could not be done if we destroyed the last known stockpiles of the Smallpox virus, the naïve intuition is correct that there exist questions that we can only answer using the virus itself.