THEORIA. Revista de Teoría, Historia y Fundamentos de la Ciencia
ISSN: 0495-4548
theoria@ehu.es
Universidad del País Vasco/Euskal Herriko Unibertsitatea
España

A. Atanasova, Nina
Validating Animal Models
THEORIA. Revista de Teoría, Historia y Fundamentos de la Ciencia, vol. 30, núm. 2, 2015, pp. 163-181
Universidad del País Vasco/Euskal Herriko Unibertsitatea
Donostia, España

Available in: http://www.redalyc.org/articulo.oa?id=339741431002

How to cite
Complete issue
More information about this article
Journal's homepage in redalyc.org

Scientific Information System
Network of Scientific Journals from Latin America, the Caribbean, Spain and Portugal
Non-profit academic project, developed under the open access initiative

Validating Animal Models*

Nina A. Atanasova

Received: 02/09/2014
Final Version: 09/04/2015

BIBLID 0495-4548(2015)30:2p.163-181
DOI: 10.1387/theoria.12761

ABSTRACT: This paper responds to a recent challenge for the validity of extrapolation of neurobiological knowledge from laboratory animals to humans. According to this challenge, experimental neurobiology, and thus neuroscience, is in a state of crisis because the knowledge produced in different laboratories hardly generalizes from one laboratory to another. Presumably, this is so because neurobiological laboratories use simplified animal models of human conditions that differ across laboratories. By contrast, I argue that maintaining a multiplicity of experimental protocols and simple models is well justified. It fosters rather than precludes the validity of extrapolation of neurobiological knowledge. The discipline is thriving.

Keywords: Animal Models, Calibration, Validity, Reliability, Experimental Neurobiology.

RESUMEN: Este artículo responde a un reto reciente: la validez de extrapolar el conocimiento neurobiológico del laboratorio animal a los humanos. Según esta objeción la neurobiología experimental, y por ende la neurociencia, está en crisis porque el conocimiento obtenido en los laboratorios de neurobiología utiliza modelos animales simplificados de las condiciones humanas que difieren de unos laboratorios a otros. Por el contrario, sostengo que mantener una diversidad de protocolos experimentales y de modelos simples está sobradamente justificado. Favorece, en vez de impedir, la validez de la extrapolación de conocimiento neurobiológico. La neurobiología es una disciplina floreciente.

Palabras clave: modelos animales, calibración, validez, fiabilidad, neurobiología experimental.

1. Introduction

Experimentation in neurobiology is a special kind of biomedical experimentation. As such, it faces the challenges commonly raised against the experimental practice of using animals for biomedical models of human conditions. Among these is the challenge for the validity of extrapolation of laboratory knowledge to phenomena that occur naturally in the world outside the laboratory. Variations of this challenge have been raised on various occasions

* The ideas developed in this paper were refined in inspiring discussions with Robyn Amos-Kroohs, John Bickle, Carl Craver, Lilia Gurova, Angela Potochnik, Bob Richardson, Rob Skipper, and Chip Vorhees. The paper was also improved in response to the comments of the audiences of the conference in Philosophy of Scientific Experimentation 4, 11-12 April 2014, Pittsburgh, PA, USA and at the Department of Philosophy and Ethics, Eindhoven University of Technology, Eindhoven, the Netherlands where I presented it on 26 August 2014. I am also grateful for extremely helpful suggestions by the two anonymous referees from THEORIA. The research reported in this paper was largely funded by Taft Research Center, University of Cincinnati with a Charles Phelps Taft Dissertation Fellowship.

for the validity of knowledge claims about the biological world produced in biological and biomedical laboratories (e.g. Diamond 1986, LaFollette and Shanks 1995, and Shanks et al. 2009). Essentially, the challenge consists in pointing out that the peculiarities and complexities of natural phenomena cannot be reproduced in the artificial context of the laboratory, especially when the knowledge claims are intended to generalize from laboratory animals to humans in the natural world. Thus, whatever knowledge can be produced on the basis of laboratory experimentation, it simply cannot capture the phenomena that occur in the natural world. Responses to these challenges include Skipper (2004), Steel (2008), and Degeling and Johnson (2013).

A version of the challenge was recently articulated for the case of neurobiological experimentation. Sullivan (2007, 2009, 2010) argues that contemporary experimental neurobiology is in a state of crisis. In her account, this is the result of the experimenters' strong commitment to strengthening the reproducibility, or reliability, of the experimental effects they study at the expense of the generalizability, or validity, of the knowledge claims produced in the laboratory. Sullivan cautions that the experimental practices of contemporary neurobiology may lead to the production of knowledge which is valid only in the local context of the laboratory in which it is produced. In her account, this is feasible because the laboratory contexts in which experimental animals are raised and tested are too controlled in order to ensure reliability of experimental procedures. Thus, they are too simplified and do not nearly match the complexity of the contexts in which the targeted phenomena occur naturally. In Sullivan's account thus, simplicity is prerequisite for reliability and complexity is prerequisite for validity. That is to say, the two norms of experimental design are in an inherent conflict with one another. Moreover, different laboratories use multiple different experimental protocols. This, according to Sullivan, prevents applying the knowledge produced in one laboratory to the effects studied in another. The lack of generalizability of knowledge across laboratories as well as knowledge produced in the laboratory to phenomena in the world outside the laboratory defines the state of crisis which Sullivan attributes to contemporary experimental neurobiology.

In other words, neurobiology is in a state of crisis because the knowledge it produces cannot be extrapolated to the phenomena it aims to study. This further precludes the integration of knowledge in neuroscience given that neurobiology is an integral part of neuroscience. This result is unsettling for both neuroscience and philosophy of neuroscience, especially given that Sullivan's analysis was the only systematic account of experimentation in neuroscience until Silva, Landreth, and Bickle (2014) recently came out of press. It explicitly challenges leading philosophical accounts of neuroscience such as Bickle (2006), Craver (2007), and Bechtel (2008) among others. Thus, it is important for philosophy of neuroscience to address the challenge for extrapolation of experimental knowledge in neurobiology raised by Sullivan.

For this reason, my purpose here is to provide an alternative account of experimentation in neurobiology such that it will make sense of the contemporary experimental practices and methodological decisions in the field. I will argue that maintaining a multiplicity of different experimental protocols and strengthening the reliability of multiple simple protocols are well justified and, contrary to Sullivan's analysis, the presence of both indicates thriving rather than crisis of experimental neurobiology.

In order to achieve my goal, I will reconceptualize the notion of validity in order to show that, contrary to Sullivan's account, maintaining a multiplicity of different ex-

perimental designs and protocols does not prevent neurobiology from producing valid knowledge claims as well as that strengthening reliability of experimental designs does not prevent the production of valid knowledge claims. I will do that by studying the explicit methodological discussions in which practicing neurobiologists engage and will analyze the practices they employ when establishing the validity of their experimental results. On this basis, I will show that neurobiologists use an elaborate calibration strategy that allows them to compare the validity of multiple simple experimental designs and this is the same strategy they use when they establish the reliability of their experimental designs. Thus, the purported tension between validity and reliability dissolves. Moreover, the strategy relies on the use and comparison of multiple experimental designs and protocols. This ultimately means that the multiplicity of experimental protocols and the stress on their reliability in neurobiology indicate thriving rather than a crisis of the discipline.

I will proceed as follows. First, I will reconstruct Sullivan's analysis of the state of contemporary experimental neurobiology. Next, I will show that the notion of validity she operates on does not adequately capture the notions of validity which are commonly in play in neurobiology. Moreover, Sullivan focuses on the validity of knowledge claims and largely ignores the practices of validating the models used in neurobiological experiments to generate those knowledge claims. Because those models are almost exclusively animal models, I articulate an account of animal models as experimental tools in neurobiology. A major step in validating the results of experiments that use animal models is to first show that the models are valid as representations of the human conditions under study. Then, I show that this validation proceeds on the basis of calibration strategy similar to the calibration used in physics (Franklin 1997) and evolutionary biology (Skipper 2004). Next, I show that the same calibration strategy is used in establishing the reliability of animal models as experimental tools. This allows me to conclude that considerations for both reliability and validity imply the use of multiple simple experimental designs and protocols. The prescriptions of reliability thus do not go against the prescriptions of validity. Moreover, using multiple experimental designs and protocols is in fact prerequisite for both validity and reliability. This ultimately shows that what Sullivan takes to indicate a crisis in experimental neurobiology is in fact beneficial for the field.

2. *The Crisis of Experimental Neurobiology*

Sullivan's analysis of the practice of experimental neurobiology makes explicit two fundamental norms of experimental design that, in her view, are in tension with one another: *reliability* and *validity*. In her account, *reliability* is a feature of the data production process which is associated with the repeatability and reproducibility of experimental effects and data. Sullivan takes this process to include experimental design and its implementation across multiple experimental trials in a given laboratory (Sullivan 2007, 7; 2009, 533). The value "reliable" is ascribed to a complete data production process when "it results in statistically analyzed data that can be used to discriminate one hypothesis from a set of competing hypotheses about an effect produced in the laboratory" (2007, 17; 2009, 534). In Sullivan's account, reliability is associated with repeatability because repetition of an experiment provides insights on potential errors as well as "some degree of certainty that an effect pro-

duced in the laboratory is not an artifact but rather the direct result of a relevant experimental manipulation" (2007, 7). Sullivan points out that experimental design begins with a process of modification of the phenomenon of interest so that it can be "translated" into an effect that can be reproduced in the laboratory. This process involves the elimination of "factors that impede the discovery of the variables that contribute to its production" (2007, 67). She notes that, "If an investigator wants to describe some feature of the effect of interest or to isolate what causes it, she necessarily has to direct the question at an effect that can be accompanied feasibly by a definitive set of competing hypotheses or claims about the effect and what produces it" (*id.*).

This process of translation, thus, involves simplification of the original phenomenon so it can be reproduced reliably in the laboratory. Simplification allows for better control of the causal variables that could have impact on the studied effect. It also enables specifying a definitive set of causal hypotheses that could account for the studied effect. Thus, reliability prescribes simplicity of experimental designs.

In addition to reliability, *validity* is another important constraint on the experimental process in neurobiology. In Sullivan's account, validity is a feature of interpretative claims produced on the basis of experimental results. It is the value ascribed to knowledge claims that can legitimately be applied to phenomena occurring in contexts different than the contexts in which they were originally produced. According to her, valid knowledge claims should be (1) the product of a reliable experimental process and (2) arrived at in laboratory contexts that are relevantly similar to the natural world contexts in which the phenomena of interest occur (2007, 88; 2009, 535). Reliability is thus necessary for validity but because of the second condition, in Sullivan's account, validity has to be considered separately from reliability. Further, because the natural contexts in which the studied phenomena occur are far more complex than the controlled environments in which laboratory animals are raised and tested, increasing the similarity between the two depends on making the laboratory contexts more complex in order to match the complexity of the natural environments they are meant to recreate. Thus, validity prescribes complexity of experimental designs.

In Sullivan's analysis, reliability and validity turn out to give opposing prescriptions for experimental design. On the one hand, reliability prescribes simplifying experimental designs in order to make it easier to detect and measure the studied effect. This would secure local, within laboratory, replication and tractability of effects. Validity, on the other hand, prescribes increasing the complexity of experimental designs in order to secure the appropriate degree of similarity between the laboratory and the natural environments of the tested subjects. Thus, she concludes that there is an inherent tradeoff between reliability and validity as constraints on experimental design. The more reliable a given experiment, the less valid the knowledge claims produced on its basis would be and vice versa. Because validity requires data produced through a reliable process it can never be increased to an extent as to lose reliability without losing validity as well. However, in Sullivan's account, reliability can be increased to an extent to which all validity can be lost (2007, 138-9; 2009, 535). Such would be the case with experiments that are so precise and controlled that they would not generate knowledge that could be extended to any context other than the one in which it was produced, including the natural world as well as the contexts of other laboratories that aim to study the same effects in different environments.

Furthermore, Sullivan reaches the conclusion that contemporary experimental neurobiology is —to put it mildly— very close to being in the situation where individual labora-

tories produce locally reliable knowledge which does not extend either to knowledge about phenomena in the world or to effects produced and studied in other laboratories. This state amounts to a crisis in the field of experimental neurobiology. In Sullivan's words,

> ...neurobiologists often make global claims about the cellular and molecular mechanisms of learning and memory. Yet, in most if not all cases of experimentation in this area, these interpretative claims are at best only *internally valid* claims – i.e., not extendable beyond the context of that laboratory that gave rise to them. This is because experimentalists in cellular and molecular neurobiology are more concerned with securing the reliability of their experiments than ensuring their validity. This makes sense, given the complexity of the brain and the complexity of learning phenomena. It is also rational, given that reliability is requisite for validity. And yet, the global explanatory aims of neurobiology, namely, to understand the cellular and molecular mechanisms of learning and memory across organisms, are not being realized. I take this to be the explanatory crisis in neurobiology (2007, 142).

Sullivan (2007) proposes a solution to the problem for the generalizability of locally valid laboratory knowledge to phenomena in the natural world. She advocates a strategy for increasing the validity of neurobiological knowledge which involves designing experiments in such a way as to gradually increase the validity of the knowledge they inspire. Her assumption is that this increase of validity can be achieved by making experimental environments and their features resemble more closely the natural environments of the phenomena they target to capture. The main strategy is to gradually increase the complexity of the experimental environments and stimuli so as to make them more similar to complex natural-world environments and stimuli which the tested organisms would encounter outside the laboratory. This can take the form of designing "incremental experiments". They would proceed by adding features (parameters) to a given experimental environment and/or stimuli so that the whole system gets to resemble more closely the targeted natural phenomenon and its environment and/or the stimuli which trigger the studied effect.

Sullivan proposes two variations on this strategy: (1) designing a series of experiments with variation of a given parameter and integration of the results of the entire series and (2) designing a series of experiments with additive variation of parameters. In the first case, the data produced in one experimental setup are compared to the data produced via a setup that differs with respect to a given feature from the first setup. The series may involve experiments that include multiple variations of a given feature. The results of the whole series are integrated and the conclusions reached this way have a broader scope than each individual experiment's results. This procedure preserves reliability but it also increases the similarity of the integrated model of the experimental system to the targeted system and thus increases validity as well. The second variation of the solution Sullivan proposes involves gradual adding of new features to a sequence of experiments in order to increase the validity of the claims about the experimental results by making the whole system more similar to the targeted natural system (2007, 160-65).

The problem with this solution is that it suggests some sort of standardization and unification of experimental procedures and designs across experiments and laboratories. It presumes that different laboratories will have to adopt unified basic procedures in order to secure the variation only of that single parameter that is supposed to be added to the basic system which was used in a prior experiment in a different laboratory. However, neurobiologists value the variety of modifications on the same basic experimental setups and pro-

cedures (Wahlsten 2001) and they have good reasons for that as well. As Würbel (2000) points out, the reproducibility of effects that can be achieved through standardization goes against the capability of experiments to produce new knowledge. It may also lead to the reproducibility of artifacts and errors. That is why Würbel concludes that "systematic variation of situations should form an integral part of all animal experimentation" (2000, 263).

3. Towards an Alternative Solution

Because of the unsettling implications of Sullivan's solution, I set out to articulate an alternative account of the practice of experimental neurobiology which is to make sense of the methodological decisions made in the field, in particular the decisions to maintain a multiplicity of experimental designs and protocols and to prefer simple reliable designs. My goal is to show that these decisions do not prevent the production of valid knowledge which extends to knowledge about effects produced in different laboratories and ultimately to phenomena that occur naturally in the world outside the laboratory.

This requires reevaluation of Sullivan's account of the tradeoff between reliability and validity. In her analysis, the two are opposed because reliability prescribes simplifying experimental protocols and procedures, which guarantees reproducibility of laboratory effects, while validity prescribes making experimental designs more complex in order to secure greater resemblance between laboratory models and their targeted natural phenomena. This means that in order to reconcile the two opposing prescriptions the association of reliability with simplicity, on the one hand, and/or the association of validity with complexity, on the other, have to be reconsidered. The association of reliability with simplicity is fairly unproblematic. It is, after all, easier to reproduce the effects of a smaller number of contributing variables. This leaves me with the option to reconceptualize validity in a way that will disassociate it from complexity.

The disassociation of validity from complexity requires reevaluation of Sullivan's analysis of the validity estimations required in experimental neurobiology. Her view is that neurobiological knowledge claims are valid, that is legitimately generalizable, when they are produced in environments that are sufficiently similar to the environments characteristic of the tested organisms in the natural world. Thus, when neurobiologists study learning and memory using laboratory rats, according to Sullivan, they have to make the laboratory environments in which they raise and test them sufficiently similar to the environments inhabited by wild rats. Because the environments inhabited by wild rats are way more complex than the typical laboratory environment, the laboratory environments have to become more complex.

Sullivan's assumption that this would ground the extension of laboratory knowledge produced via animal experimentation to humans in the natural world is highly problematic. She admits that most of experimental neurobiology targets cognitive processes that occur naturally in humans not in wild animals. Nevertheless, she stipulates that making laboratory rats resemble their natural-world counterparts will make them more similar to humans and thus will ensure the validity of extrapolation of laboratory knowledge produced in animal studies to humans (Sullivan 2007, 153-4). Her strategy to increase the requisite similarity would be adequate in studies that aim to extrapolate laboratory knowledge produced in one species to members of the same species outside of the laboratory. In other

words, Sullivan's analysis of validity would be adequate for cognitive ethology (e.g. when knowledge about learning in laboratory rats is generalized to knowledge about rat learning in general) or for psychological experiments on human subjects which aim to extrapolate knowledge about humans in the laboratory to humans in the natural world.

This problematic assumption made by Sullivan is perhaps due to her focusing almost exclusively on notions of validity and reliability employed in psychology. However, spelling out an adequate account of neurobiological animal experimentation would be better grounded if it incorporated considerations of the reliability and validity criteria employed in neurobiological experimentation broadly construed, including neurobiology of cognition and emotions as well as medical neurobiology of cognitive, psychiatric, neurological, and mood disorders. Studying the discussions of modeling in biology, especially the discussions about the epistemological features of model organisms and animal models, which are basic tools of experimental biology, would provide important insights into the nature of experimentation in neurobiology as well. This is important especially given that historically neurobiological animal experimentation has been developed in biology departments and medical schools more often than in psychology departments (Haug and Whalen 1999).

3.1. Animal Models as Experimental Tools

Experimental neurobiology is a biological discipline and thus its philosophical account should be situated within philosophy of biology, and more precisely on the intersection of philosophy of biology and philosophy of psychology. In their experimental practices, the biological sciences make an extensive use of material models that include animal organisms or their parts. The most widely utilized tools for biomedical and neurobiological experimentation are animal models. Animal models are systems of material objects containing live modified or intact non-human animals or their parts. They are used for (very often) invasive experimentation that relies on inducing or simulating pathological physiological conditions in the experimental subjects. The ultimate goal of this kind of experimentation is to learn about the occurrence, development, and treatment of the conditions of interest, where the interventions used to produce the studied effects (for the most part) would not be morally permissible on human subjects. The animals in such models are typically used as proxies or substitutes for humans. Animal models in neurobiological experimentation share many common features with biomedical animal models and other related material biological models such as model and experimental organisms (Ankeny and Leonelli 2011) all of which contain animal organisms or their tissues or organs as their integral parts. Unlike other types of models used in science, models containing animals as substitutes for humans are supposed to be similar to the phenomena they target to represent. This is so because the extrapolation of knowledge about laboratory animals to naturally occurring human processes and conditions relies on analogical arguments and analogical arguments require establishing relevant similarities and/or identities between the two analogues which are compared.

Hesse (1966/1963) grounds this sort of inference on the basis of material analogy. She describes it as comparison between two analogues defined in terms of their properties and the relationships between them. An analogy includes horizontal and vertical relationships between those properties. The horizontal relationships are the relationships of similarity and dissimilarity between the properties of the two analogues. The vertical relationships

are the causal relationships that hold between the properties of each analogue. The inference based on analogy proceeds as follows: if two objects can be shown to share identical or very similar sets of horizontal properties, then whatever is known to be a causal relationship between those properties of one of the objects can be assumed to be present in the other object as well. For example, humans and rats share similarities of their nervous systems in terms of corresponding brain structures, neurotransmitters, etc. They also share similarities in their behaviors. For example, both humans and rats are capable of learning to locate the position of an object of interest provided that it remains in the same location during multiple training situations. Because we know that the removal of the hippocampus in humans leads to inability to form long-term declarative memories we can infer by analogy that similar deficits will occur in rats provided that their hippocampus is removed. Analogies are used when one of the analogues is familiar and its internal properties and their causal relationships are well known. This analogue is then used as a model on the basis of which knowledge about some unknown features of the other analogue can be inferred. Animal models thus serve as such models for human conditions.

The arguments that justify extrapolation of knowledge about laboratory animals to humans in the natural world depend on this sort of material analogy. All biological models that include live organisms or their tissues fall under the category of material models. Material models are central to theoretical and experimental biology (Laubichler and Müller 2007, Leonelli 2007). All models that include animals as their parts rely on the assumption that there are relevant similarities and commonalities between all animal species, or at least between the species involved in any given comparative study. Thus, by analogy, what is known about one species can be extended as knowledge about the relevantly similar other species. The argument for using certain species for drawing conclusions about the biological makeup of others is thus an argument from analogy. One of its premises asserts the relevant similarities between the two species (horizontal relation of similarity). This is taken to justify the assumption that if something is experimentally learned about a given species (i.e. some vertical causal relation is established in the model) then it probably applies to other species of interest (i.e. the same causal relationship holds between the corresponding properties of the analogue). Another premise asserts that something is known about the experimental animals. The conclusion of the argument asserts that what is known about the experimental animals applies to the targeted animals as well.

In the case of mood or neurological disorders, the relevant similarities between the model and its target are observable behaviors and corresponding physiological states. When animals are used to study depression and schizophrenia, for example, their behaviors are taken to correspond to the analogous human behaviors associated with the studied conditions. In order to be able to extrapolate knowledge from the animal models of these conditions, neurobiologists have to first show that what is known about the human conditions, or at least something very similar to them, can be reproduced or simulated in animal organisms. This is the process of model building in which an animal model is designed and developed so that it can simulate the studied condition, e.g. depression (Willner 1991a).

In the process of designing an animal model of a given condition, the human condition is initially a model for its animal model in a sense very similar to that articulated by Hesse. At this stage the available knowledge about the human condition is greater than the knowledge available about the animal model. Only after the human condition has been successfully modeled, or simulated, in the animal model can it be used as an instru-

ment to generate further knowledge that can then be extrapolated back to humans. The philosophical discussions about the validity of extrapolation of knowledge about animal models to human conditions have largely focused on this second stage. However, what makes the extrapolation legitimate actually happens at the first stage, the stage of model building.

3.2. Calibration of Animal Models

The stage of model building is characterized by considerations for getting the model to match its target in relevant respects. It is very well captured by the notion of calibration of laboratory models to features of their targets used in Skipper (2004). Skipper employs this notion of calibration to capture the justification of the extrapolation of results obtained in laboratory models of populations of organisms to evolutionary processes occurring in natural populations. He shows that extrapolation is justified when laboratory models —populations of organisms in the case he studies— "are matched, or tuned, or *calibrated,* to relevant features of natural populations" (2004, 372). Similarly, animal models are tuned to exhibit features and behaviors similar to the ones characteristic of humans in the studied conditions.

Skipper (2004) develops his notion of calibration of biological laboratory models as an extension to the notion of calibration articulated in Franklin (1997). Franklin (1997) defines *calibration* as "the use of surrogate signal to standardize an instrument", which "is an important strategy for the establishment of the validity of experimental results" (Franklin 1997, 31). He further distinguishes calibration from measurement. While in measurement the results of an experiment are unknown, in calibration, the expected results are known in advance. Calibration aims at getting an instrument to reproduce known effects. Franklin also introduces an extended version of the calibration strategy, which includes estimating the validity of an experimental result by examining the analysis procedures employed in its formulation (Franklin 1997, 75). The analysis of some typical examples of establishing validity of animal models articulated in the next section shows that the notion of calibration can be extended to capture the experimental practices employed for the validation of animal models in neurobiology.

4. Establishing Validity

The types of validity of animal models explicitly addressed by neurobiologists typically include predictive validity, face validity, construct validity, and criterion validity (Vorhees 1987, Willner 1991b, Sufka et al. 2009, Warnick et al. 2009, Hymel et al. 2010). *Predictive validity* is often assessed by the responses of animal models to drugs whose effects are known. For example if a model of anxiety exhibits lack of anxious behavior when treated with anxyolitics (drugs known to be effective in the treatment of human anxiety), then it is considered to have predictive validity. If it also responds in the expected way to chemicals that are known to worsen the condition the model has even better predictive validity. Predictive validity, according to Vorhees, is "the ability of a test to predict effects from an incomplete or partial data set" (Vorhees 1987, 458). In the case of preclinical toxicology tests, it is important to show that an animal test will be predictive for the effects that a given

chemical would produce in humans. It is crucial to establish a test's "cross-species predictive validity". The strategy here is to test animals' reactions to substances that are known to be toxic for humans but it is also important for the model to be capable of correctly identifying substances that are non-toxic for humans. When an animal model exhibits effects similar to those known to occur in humans in reaction to multiple substances, it can be expected to be successful in predicting the effects of other substances on humans based on the animal reactions produced in the experimental system.

Face validity "refers to a phenomenological similarity between the model and the disorder it simulates: on one hand, the model should resemble the disorder, while on the other, there should be no major dissimilarities" (Willner 1991b, 9). It is usually established by inducing in the test organism behaviors that resemble the target behaviors superficially. The requirement is to make the model look like the target phenomenon "on the face of it" regardless of the underlying mechanisms that are causally responsible for what is observed.

Evaluating the theoretical rationale behind a given animal model is what is involved in estimating its *construct validity*. Construct validity is based not only on comparison of behaviors or superficial similarities. It also requires proper mapping between the hypothesized mechanisms underlying the human and the animal behaviors respectively. As Sullivan herself properly identifies this sort of validity, it requires establishing the identity of the hypothesized causally responsible constructs and whatever the experiment actually measures.

Criterion validity refers to the "ability of a test to measure a characteristic that can be independently defined" (Vorhees 1987, 457). Vorhees points out that this approach to establishing validity in toxicology can be successful if a database with known neurotoxic agents is made available. A test is then estimated for its efficacy when applied to known neurotoxins.

All of the conceptualizations of validity that can be identified as operative in experimental neurobiology require the analysis of the results of multiple tests and experiments. If this analysis shows that the results converge or at least are compatible with one another and they properly exhibit the target effects, the tested animal models are considered valid representations, or simulations, of the studied phenomena. As such the notion of convergence typically underlies all of the notions of validity operant in neurobiological animal model experimentation.

4.1. Convergent Validity

Campbell and Fiske (1959) discuss a validation procedure which requires the convergence of measurements from independent tests and techniques. They note that this requirement underlies most notions of validity in experimental psychology. Because all experimental procedures are potentially fallible, a multiplicity of independent experimental techniques which converge in the results they produce will ensure the validity of the knowledge generated this way. Campbell and Fiske term this type of validity *convergent validity*. It captures well the commonalities shared by the different types of validity with which neurobiologists operate.

In neurobiology, testing for convergence may take different forms. It may involve testing the model for reproduction of target effects using different factors whose effects are known in the target system. For example, an animal model developed for drug testing is first tested against multiple known substances whose effects are well documented in humans in order to check whether the animal responds in similar ways to the known treatments. It may in-

volve using multiple different tests to check whether the model captures multiple aspects of the target phenomenon as well as to check the tests against each other in order to avoid conflation of artifacts of the experimental setup with genuine targeted effects. In the following subsections, I provide examples of how these validation goals are actually implemented.

4.2. Testing for Convergence against Multiple Known Factors

In their experiment performed with the goal to establish the validity of the fowl chick model as a simulation of the anxiety-depression continuum, Warnick et al. (2009) tested their model for reproducing the target phenomenon. The model is designed to mimic the anxiety-depression continuum symptoms by inducing —through social separation— behavior which is taken to correspond to the human behavior characteristic of the syndrome. In this experiment, Warnick and colleagues aimed to strengthen the validity of the anxiety-depression continuum chick model. According to the experimental paradigm they had previously developed, socially raised chicks exhibit panic-like anxiety induced by social separation for the first 5 min of a 2-hour-long test. The chicks' anxiety is measured on the basis of the rate of their distress vocalizations which drop in half within 20-25 minutes after the beginning of the test and remain relatively steady for the remainder of the 2-hour-period. The latter phase of the test is taken to mimic the human state of depression conceived as learned helplessness (Warnick et al. 2009, 144).

For their new experiment, they used 5 to 7-day old socially raised *Gallus gallus* cockerels. Different groups of the animals were treated with 2 chemically different medically approved anxiolytics and 3 different known antidepressants. The animals were subjected to the behavioral paradigm described earlier. For each drug, Warnick and colleagues compared the effects of multiple concentrations with the effects of treatment with an inactive substance in socially isolated animals and animals put in cages with two conspecifics. As expected, the social animals did not exhibit significant distress symptoms whereas the isolated animals exhibited high rates of distress vocalization in the phase of anxiety and the typical drop in half of the rate of distress vocalizations during the depression-like stage. All of the tested drugs, except one of the antidepressants led to reduced rates of distress vocalizations during the anxiety phase indicating the expected anxiolytic effect. The only antidepressant that did not induce such effect was interpreted as a true negative because it had been previously shown to be ineffective in the treatment of situationally bound panic disorder (2009, 153). As expected, in the depression-like phase, all tested antidepressants led to increased rate of distress vocalizations. This was taken to indicate antidepressant action. These results, combined with further analysis of depression-biomarkers in the blood of tested animals and the results of previous animal and clinical studies, were taken to strengthen the validity of the "chick anxiety-depression continuum as a clinical simulation and putative preclinical screening assay" (2009, 153).

The establishment of the validity of the chick model relied on the convergent results from multiple drug tests which reproduced —in animals— effects that were known to occur in humans. Warnick and colleagues argued that because the validity of the animal model simulation was established in this experiment, the same behavioral paradigm could be used for testing other treatments as well. In other words, it could be used as an experimental tool that would produce valid results. The knowledge generated this way could then be extrapolated to humans.

4.3. Convergent and Compatible Results from Different Tests

Another way to guarantee the validity of a model of a given process is to embed tests within more complex experimental arrangements. Such is the purpose of using test batteries. Test batteries are sets of multiple experimental tests that are commonly employed in the study of the complex functions, or dysfunctions, of the nervous system. Vorhees (1996) defines a test battery for assessing central nervous system (CNS) function as "a collection of tests, assembled to provide a systematic assessment of the neurological, cognitive, or emotional status of the animal [laboratory rodents in the case analyzed by Vorhees]" (227). The tests employed in different batteries vary according to the different experimental goals of the studies that employ them. Vorhees discusses toxicological test batteries and distinguishes between broad spectrum, focused, and mixed test batteries. Broad spectrum test batteries can be comprehensive, "designed to provide broad profile of all major functional domains" or screening, "intended to detect only major neurobehavioral deficits" (228). Broad spectrum test batteries include multiple tests among which locomotor activity, auditory startle habituation, operant conditioning, etc. They are often used to assess the spectrum of the effects on multiple CNS functions produced by compounds with unknown neurotoxicity. Focused test batteries are employed in experiments that explore a particular hypothesis. Vorhees points out that among "the most common types of focused batteries are those investigating the effects of particular drugs or surgical treatments on different types of learning" (*id.*). Mixed test batteries share characteristics with the other two types. Like the broad spectrum test batteries, they are used in multi-functional analysis of unknown effects. However, they are employed when there are strong reasons to believe that a given treatment or other causal agent has a particular effect on the functions of the central nervous system. In this sense, mixed batteries are more similar to the focused test batteries because there is a particular hypothesis about a causal action that is tested in the experiment. Vorhees summarizes the main reasons that motivate the design and use of test batteries as follows,

> ...multiple tests provide a comprehensive assessment of different CNS functional domains. Multiple dependent measures provide converging data on each functional domain by measuring the underlying variable with different degrees of overlap and also different degrees of error. Different ways of measuring the same variable provide different types of measurement error, the sum of which may be expected to cancel each other out, leaving the best possible assessment of the central process of interest. For example, learning cannot be measured directly and must rely upon sensory and motor capacities, but differences in vision, hearing, motivation, motoric ability, and other factors can affect performance on tests of learning. By measuring different performance factors, one can generally be assured that treatment-related differences between groups are attributable to differences in learning rather than other factors (1996, 229).

Vorhees provides an argument for the epistemic utility of constructing test batteries with at least partial functional overlap. When one and the same function is assessed with multiple tests this "overlap provides convergence of evidence thereby assuring that if an effect is found on, for example, memory, the effect is real, reliable, and not an artifact of performance factors" (Vorhees 1996, 230).

As in the previous example, test batteries rely on some sort of converging of results produced on the basis of multiple tests or different experimental arrangements. What is im-

portant is that producing compatible and converging results on the basis of multiple tests or experimental arrangements strengthens the validity of each line of converging results. These results validate one another.

Both examples of establishing convergent validity can be considered as applications of the calibration strategy in a broad sense. The validation of the chick anxiety-depression continuum model uses calibration as testing against multiple known factors to confirm that their known effects are reproduced in the animal model. The strategy Warnick et al. used is presenting a surrogate signal to check whether the model reproduces the expected results. The chemicals applied in the testing of the model were used as surrogates for the novel compounds which would eventually be tested in the model if it was established as a device that would detect their effects. Test batteries exemplify calibration that employs testing different experimental setups against each other to check whether they produce converging and/or compatible results. The results of the individual tests within a battery are tested against each other to ensure that the apparatus detects the phenomena it is supposed to detect. This sort of analysis is in line with the extended version of calibration as a strategy for justifying the validity of experimental results. It also makes explicit the need for multiple experimental designs, protocols, and procedures which produce convergent (or at least compatible) results in order to establish the validity of the knowledge claims produced on the basis of animal experimentation. Contrary to Sullivan's view, animal models get validated not in spite of but by virtue of multiple different experimental protocols. Neurobiologists claim their animal models to be valid when the results they produce are compatible, consistent, and/or convergent with the results produced in different experimental setups and/or in variations of the same experimental setup. So far, I have addressed Sullivan's concern about the lack of standard and unified experimental procedures in neurobiology. In the next section, I address her concern about the opposing prescriptions that validity and reliability impose on experimental design.

5. Establishing Reliability

Calibration is also employed for establishing the reliability of animal models in neurobiology when identical or slightly modified animal model designs are tested for reproducibility of effects and convergence of results in different laboratories. The reliability established this way is broader than the notion of reliability Sullivan opposes to validity. On her account, neurobiologists aim at strengthening the reliability of the procedures they use within their laboratory. That is to say, they focus on strengthening the *intra-laboratory* reliability of their experimental systems. However, neurobiologists are sometimes interested in establishing the *inter-laboratory* reliability of their models. This interest is justified given that good science requires experimental results to be capable of being replicated in a variety of circumstances, including other laboratories. The following subsection provides an example of how neurobiologists establish both types of reliability.

5.1. Testing for Reproducibility in Different Laboratories

Vorhees (1987) discusses the implications of a collaborative study in which the results from two experiments involving identical behavioral tests (employing identical appara-

tuses, strains of animals, and protocols) performed in six different American laboratories were compared. The first experiment traced the effects of prenatal exposure to amphetamine and the second studied the effects of prenatal exposure to methylmercury. There was a "high degree of comparability between all laboratories for a given study" and a "remarkable degree of comparability within laboratories" from the first to the second experiment (Vorhees 1987, 450). In other words, the study showed that the tested battery had both inter- and intra-laboratory reliability. It did so even though there were differences in the baseline results between laboratories. Despite those differences, the detection power of the tests was similar across different laboratories. This meant that the tested model's behavior was reproduced reliably. These results were compared to the results of a German laboratory whose study used identical apparatus for activity testing but modified protocols. The results were also comparable and because an "excellent degree of between laboratory comparability was obtained", Vorhees concluded that this "refutes the notion that almost any procedural variation has major effects on measures of behavior" (451).

The study demonstrated that in order for reliability to be established, laboratories have to keep a record of a baseline of controls which can be compared with a contemporaneous control group for establishing a baseline for any experiment. This would provide means for estimating intra-laboratory reliability. Establishing a baseline is also crucial for the comparability of test results between laboratories. This is necessary for estimating inter-laboratory reliability. Establishing a baseline for the results of behavioral tests and comparing baseline to group differences between laboratories takes the form of calibrating animal models. The results of different laboratories are tested for convergence against each other. The expected effect is that all laboratories would produce converging results. Once the convergence of results is established, the procedures employed in these studies can be used as a standard against which future experimental designs can be calibrated.

5.2. Strengthening Reliability while Strengthening Validity

It is now evident that the same sort of strategy is used for establishing both the reliability and the validity of animal models in neurobiology. This practice is well captured by the notion of *robustness analysis* articulated by Wimsatt (1981). Robustness analysis, according to Wimsatt, is, among other things, the "use of multiple means of determination" with the goal to confirm the validity, or reality, of a given postulated theoretical construct, the reliability of a test or instrument, as well as for calibration or recalibration of measuring devices (Wimsatt 1981, 63). However, robustness analysis is a much broader family of techniques and strategies than the calibration strategy discussed in the previous sections. Philosophers of neuroscience have tended to focus on its usefulness in determining the reality of unobservable entities and/or their properties (e.g. Craver 2007, Eronen 2012).

My purpose here is to articulate the practice for ensuring reliability of experimental tools and the validity of the knowledge claims generated with their help. This is why I need to restrain my analysis to a narrower domain of application. Campbell and Fiske's account is already among Wimsatt's inspirations and it is focused on the relationship between reliability and validity. This motivates my choice not to employ robustness analysis even though it is an already established notion in philosophy of neuroscience.

Campbell and Fiske's focused analysis articulates quite clearly the commonalities and differences between validity and reliability. In their account, "Validation is typically *conver-*

gent, a confirmation by independent measurement procedures" (Campbell and Fiske 1959, 81). Thus, it requires multiple tests and procedures in order to be established. Similarly, reliability also requires reproducibility over multiple tests and repetition of procedures. However, while reliability requires reproducibility, or convergence, of results produced with identical tests, validity requires reproducibility, or convergence, of results produced with different tests and/or different causal factors. As Campbell and Fiske put it,

> Both reliability and validity concepts require that agreement between measures be demonstrated. A common denominator which most validity concepts share in contradistinction to reliability is that this agreement represents the convergence of independent approaches. [...] Independence is, of course, a matter of degree and, in this sense, reliability and validity can be seen as regions on a continuum. [...] Reliability is the agreement between two efforts to measure the same trait through maximally similar methods. Validity is represented in the agreement between two attempts to measure the same trait through maximally different methods (1959, 83).

Convergent validity is achieved when measurements produced using different techniques converge. Even though Campbell and Fiske represent reliability and validity as located on different ends of a spectrum, this does not necessarily put them in opposition, as does Sullivan's account. An animal model can be calibrated to reproduce reliably the targeted effects. It may produce results which converge towards the results produced using modified protocols and procedures (as in the case described by Vorhees 1987) or towards the results of independent tests and techniques (as in the case of test batteries). The calibration strategy can be used in both cases. Campbell and Fiske's notion of validity as requiring the convergence of multiple tests, thus allows for the same procedures that strengthen reliability to strengthen validity of the neurobiological knowledge produced on the basis of animal experimentation. This means that using the calibration strategy, it is possible to strengthen both reliability and validity at the same time. This also shows that focusing on strengthening the reliability of simple laboratory models is consistent with establishing the validity of their results provided that multiple laboratory models produce results that converge.

One might object that this procedure does not account for the tension between the opposing prescriptions of validity and reliability. In Sullivan's account, validity prescribes complexity and reliability prescribes simplification of experimental designs. The calibration strategy seems to favor simplicity because simpler designs will more readily reproduce the expected results.

My response is that this opposition results from two inaccurate beliefs that Sullivan holds, namely: (1) that animals in the wild are more similar to humans than laboratory animals and (2) that validity of neurobiological knowledge claims requires them to be the product of laboratory systems including organisms and their environments that match the complexity of the target natural world systems including organisms and their natural environments. The combination of (1) and (2) leads Sullivan to the conclusion that validity requires the laboratory systems to be so designed as to have the environments of the tested organisms to match the complexity of the environments that these organisms are likely to encounter in the wild. Both of these assumptions are unjustified.

1. The first is unjustified because many neurobiological animal models are such that they are not meant to be representative of the species from which the organisms

used in experiments are drawn. Rather, they are meant to represent human conditions directly. In this sense, laboratory animals may be more similar (at least in the relevant respects) to humans than they are to their counterparts in the wild. This point is well exemplified in the case study developed by Ankeny et al. (2014). Ankeny and colleagues study animal model research on alcohol addiction in North America since the mid-twentieth century. Their analysis shows that in the kind of experiments employed in the field, it is crucial to situate the experimental animals in environments that are representative of the environments in which the human targets of the models are typically situated. The goal is to get the animals to exhibit behaviors similar to the behaviors of humans in those environments (Ankeny et al. 2014, 488). Note that those environments and behaviors do not have to be typical of the animals in their natural environments. This is so because "interest in human alcoholism is what drives research interests, not scientific interest in understanding alcohol's effects on non-human animals for its own sake" (494). In this sense, the experimental animals employed in the neurobiological study of human conditions are more similar to humans than their counterparts in the world outside the laboratory. Therefore, Sullivan's concern that laboratory animals are raised and tested in conditions quite different than their natural environments does not affect the validity of extrapolation of laboratory knowledge produced through animal experimentation to knowledge about human conditions in the world outside the laboratory.

2. The second assumption is unjustified because Sullivan takes the complexity of natural environments to be the crucial characteristic that has to be present in laboratory environments in order to ensure the match between laboratory environments and the environments in which the target conditions occur naturally. However, in order to secure reproducible knowledge, laboratories have to operate with highly controlled environments. Nevertheless, this does not mean that the relevant factors from the target environments cannot be reproduced in a controlled environment. As Ankeny and colleagues point out, while in areas of biomedical research such as genetic studies, the environment is taken as a background against which the organism is studied in isolation, in behavioral studies (in their case alcohol research), "model validity is assessed with reference to the features of both the organism itself and the environment and experimental settings within which it is being studied" (2014, 487). This is to show that neurobiologists do take the role of environment to be a relevant factor in their models. However, nothing in this realization implies that the relevant feature of the natural environment to be matched in the laboratory is its complexity. The examples discussed in the previous sections show how the validity of the knowledge claims obtained in controlled settings can be established on the basis of the match, or convergence, of multiple relatively simple experimental setups. I take this to be indicative of the adequacy of the notion of convergent validity in capturing the practices of validation of animal models in experimental neurobiology.

When validity of neurobiological animal models is conceptualized as convergent validity, this allows the application of the calibration strategy for the purpose of strengthening both reliability and validity. Thus, the inherent contradiction between the prescriptions im-

posed on experimental design by reliability and validity identified by Sullivan (2007, 2009) dissolves. In effect, maintaining a multiplicity of simple experimental designs and protocols is in fact justified and shows that experimental neurobiology is not in a state of crisis and it is actually thriving.

6. Conclusion

In this paper, I aimed at providing an account of contemporary neurobiology that would make sense of the experimental practices and the methodological decisions prevalent in the field. Because experimental neurobiology relies almost exclusively on the use of animal models to study the human nervous system, I analyzed their building and justification as valid representations of human conditions. I argued that the validation of animal models in neurobiology often proceeds through their calibration in order to reproduce the targeted effects. This calibration takes the form of testing of animal models against factors whose effects in humans are well known, e.g. the effects of medically approved drugs. It often takes the form of employing multiple tests whose results are checked against each other for convergence and/or consistency, e.g. in test batteries. Calibration is also used for establishing reliability in neurobiological experiments. Because, it is the same strategy that is used to establish both validity and reliability, I argued that the tension which Sullivan identifies between validity and reliability as norms of neurobiological experimental design ultimately dissolves. In other words, maintaining a multiplicity of experimental protocols for the study of identical natural world phenomena is methodologically well justified. Furthermore, the tendency of different laboratories to strengthen the reliability of their idiosyncratic protocols does not preclude the validity of the knowledge produced in one laboratory as knowledge about the effects studied in other laboratories and ultimately about the phenomena in the world outside the laboratory. This is so because when multiple different experimental designs and protocols produce converging results they increase each other's validity.

REFERENCES

Ankeny, Rachel A. and Sabina Leonelli. 2011. What's so special about model organisms? *Studies in History and Philosophy of Science* 42: 313-23.

Ankeny, Rachel A., Sabina Leonelli, Nicole C. Nelson and Edmund Ramsden. 2014. Making Organisms Model Human Behavior: Situated Models in North-American Alcohol Research, since 1950. *Science in Context* 27 (3): 485-509.

Bechtel, William. 2008. *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. New York: Routledge.

Bickle, John. 2006. Reducing mind to molecular pathways: Explicating the reductionism implicit in current cellular and molecular neuroscience. *Synthese* 151 (3): 411-34.

Craver, Carl F. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Clarendon Press; New York: Oxford University Press.

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

Degeling, Chris and Jane Johnson. 2013. Evaluating Animal Models: Some Taxonomic Worries. *Journal of Medicine and Philosophy* 38: 91-106.

Diamond, Jared. 1986. Overview: Laboratory Experiments, Field Experiments, and Natural Experiments. In *Community Ecology*, eds. Jared Diamond and Ted Case, pp. 3-22. New York: Harper and Row Publishers.

Eronen, Markus I. 2012. Pluralistic physicalism and the causal exclusion argument. *European Journal for Philosophy of Science* 2 (2): 219-232.

Franklin, Allan 1997. Calibration. *Perspectives on Science* 5 (1): 31-80.

Haug, Marc and Richard E. Whalen. 1999. Introduction. In *Animal Models of Human Emotion and Cognition*, eds. Marc Haug and Richard E. Whalen, pp. 3-12. Washington, DC: American Psychological Association.

Hesse, Mary. 1966/1963. *Models and Analogies in Science.* University of Notre Dame Press.

Hymel, Kristen A., Amy L. Salmeto, Eun Ha Kim and Kenneth J. Sufka. 2010. Development and Validation of the Chick Anxiety-Depression Continuum Model. In *Translational Neuroscience in Animal Research*, ed. Jason E. Warnick and Allan V. Kalueff, 83-110. New York: Nova Science Publishers.

LaFollette, Hugh and Niall Shanks. 1995. Two Models of Models in Biomedical Research. *The Philosophical Quarterly* 45 (179): 141-60.

Laubichler, Manfred D. and Gerd B. Müller. 2007. Models in theoretical biology. In *Modeling Biology: Structures, Behavior, Evolution*, eds. Manfred D. Laubichler and Gerd B. Müller, pp. 1-14. Cambridge, Mass.: MIT Press.

Leonelli, Sabina. 2007. What is in a model? Combining theoretical and material models to develop intelligible theories. In *Modeling Biology: Structures, Behavior, Evolution*, eds. Manfred D. Laubichler and Gerd B. Müller, pp. 15-36. Cambridge, Mass.: MIT Press.

Shanks, Niall, Ray Greek and Jean Greek. 2009. Are animal models predictive for humans? *Philosophy, Ethics, and Humanities in Medicine* 4 (2).

Silva, Alcino, Anthony Landreth and John Bickle. 2014. *Engineering the Next Revolution in Neuroscience*. New York: Oxford University Press.

Skipper, Robert A. Jr. 2004. Calibration of Laboratory Models in Population Genetics. *Perspectives on Science* 12(4): 369-93.

Steel, Daniel P. 2008. *Across the Boundaries: Extrapolation in Biology and Social Science.* Oxford: Oxford University Press.

Sufka, Kenneth J., Morgan Weldon and Colin Allen. 2009. The Case for Animal Emotions: Modeling Neuropsychiatric Disorders. In *The Oxford Handbook of Philosophy and Neuroscience*, ed. John Bickle, 522-36. Oxford & New York: Oxford University Press.

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: 511-39.

Sullivan, Jacqueline A. 2010. Reconsidering 'spatial memory' and the Morris water maze. *Synthese* 177: 261-83.

Vorhees, Charles V. 1987. Reliability, Sensitivity and Validity of Behavioral Indices of Neurotoxicity. *Neurotoxicology and Teratology* 9: 445-64.

Vorhees, Charles V. 1996. Design Considerations in the Use of Behavioral Test Batteries for the Detection of CNS Dysfunction in Laboratory Animals. *Mental Retardation and Developmental Disabilities Research Reviews* 2: 227-33.

Wahlsten, Douglas. 2001. Standardizing tests of mouse behavior: Reasons, recommendations, and reality. *Physiology and Behavior* 73: 695-704.

Warnick, Jason E., C. J. Huang, Edmund O. Acevedo, and Kenneth J. Sufka. 2009. "Modeling the anxiety-depression continuum in chicks." *Journal of Psychopharmacology* 23: 143-56.

Willner, Paul. 1991a. Animal models as simulations of depression. *Trends in Pharmacological Sciences* 12: 131-36.

Willner, Paul. 1991b. Behavioural models in psychopharmacology. In *Behavioural models in psychopharmacology: theoretical, industrial and clinical perspectives*, ed. Paul Willner, pp. 3-18. Cambridge, New York & Melbourne: Cambridge University Press.

Wimsatt, William C. 1981. Robustness, Reliability, and Overdetermination. In *Characterizing the Robustness of Science: After the Practice Turn in Philosophy of Science*. 2012, pp. 61-87. Reprinted from *Scientific Inquiry in the Social Sciences*, eds. M. Brewer and B. Collins. 1981, pp. 124-63. San Francisco: Jossey-Bass.

Würbel, Hanno. 2000. Behaviour and the standardization fallacy. *Nature Genetics* 26: 263.

Nina Atanasova is a Lecturer in philosophy at the Department of Philosophy and Religious Studies, The University of Toledo (Ohio, USA). Her main research interests are in philosophy of neuroscience, medicine, and cognitive science, as well as neuroethics and biomedical ethics.

Address: The University of Toledo. Department of Philosophy and Religious Studies. Main Campus, University Hall, Room 4600. 2801W. Bancroft. Toledo, OH 43606, USA. E-mail: nina.atanasova@utoledo.edu