universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Methodological Limitations of Using Animal Models in Cognitive Science Research

verfasst von | submitted by

Marin Ivanov Georgiev

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 013
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Joint-Degree Middle European interdisc. master prog. in Cognitive Science
Betreut von | Supervisor:	Zsofia Viranyi PhD

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Methodological Limitations of Using Animal Models in Cognitive Science Research

verfasst von | submitted by

Marin Ivanov Georgiev BA

angestrebter akademischer Grad | in partial fulfillment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 013
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Joint-Degree Middle European interdisc. master prog. in Cognitive Science
Betreut von | Supervisor:	Priv.-Doz. Zsofia Viranyi, PhD.

Methodological Limitations of Using Animal Models in Cognitive Science Research

A Master's Thesis by

Marin Ivanov Georgiev

University of Vienna

2024

Table of Contents

1. Introduction……..…..*06*

1.1 Animal Modeling Within the Field of Cognitive Science.…….……..…..08

1.2 Reliability of Animal Models as a Method.…….…….....11

1.3 Brief History of Cognitive Animal Models.…….……..…..18

1.4 Methodological Approaches in Deploying Cognitive Animal Models.…….……..…..28

2. First Case Study – Genetic Mice Models of Obsessive-Compulsive Disorders.…….…….....*37*

2.1 Obsessive-Compulsive Disorder – Short Overview.…….……..…..38

2.2 Cognitive Science and its Relevance in Obsessive-Compulsive Disorders.…….……..…..41

2.3 Case Study Overview.…….……..…..42

2.4 First Case Study Discussion – Critical Analysis and Reflection.…….……....51

3 Second Case Study – Modeling Primates for Studying Language Evolution……..*62*

3.1 Experimental Design and Deployment of the Model.…….……....63

3.2 Second Case Study Discussion – Critical Analysis and Reflection.…….……....70

4. Discussion……..…..*79*

4.1 Ethical implications of using animal models.…….……....82

Bibliography……..……....*88*

Keywords: animal models, translational research, cognitive science, external validity, internal validity, transgenic models, behavior, extrapolation

Abstract

Animal models are frequently used in research practice with the aim to help us gain a deeper understanding of humans. This is particularly evident in pharmacological research, where animals are primarily deployed for testing drugs' safety and efficacy (Robinson et al., 2019). In cognitive science, however, animal models are deployed with a greater methodological variety to uncover the secrets of human cognition, ranging from neuroimaging and behavioral experiments, to transgenic research (Ting & Feng, 2011; Hopkins et al., 2012). Although there is already an ongoing debate about the reliability of animal experimentation in biomedical research, there has been relatively little attention paid to how animal models affect our assumptions and findings on human cognition (Akhtar, 2015). In this thesis, we aim to address the methodological gaps and limitations of animal models when the subject of interest is human cognition. For that, we performed a rigorous and critical analysis of two case studies that showcase different applications of animal models in cognitive research. Our findings highlight the fundamental translational challenges such an approach introduces, underscoring the importance for a shift towards methodologies that are both empirically and ethically robust, in order to ensure a solid and reliable understanding of human cognition.

Schlüsselwörter: Tiermodelle, Translationale Forschung, Kognitionswissenschaft, Externe Validität, Interne Validität, Verhalten, Extrapolation

Deutsche Zusammenfassung

Tiermodelle werden häufig in der Forschungspraxis eingesetzt, um ein tieferes Verständnis des Menschen zu erlangen. Dies ist insbesondere in der pharmakologischen Forschung der Fall, wo Tiere hauptsächlich zur Prüfung der Sicherheit und Wirksamkeit von Medikamenten eingesetzt werden (Robinson et al., 2019). In der Kognitionswissenschaft hingegen werden Tiermodelle mit einer größeren methodischen Vielfalt eingesetzt, um die Geheimnisse der menschlichen Kognition zu lüften, angefangen von Neuroimaging- und Verhaltensexperimenten bis hin zur transgenen Forschung (Ting & Feng, 2011; Hopkins et al., 2012). Obwohl es bereits eine laufende Debatte über die Zuverlässigkeit von Tierversuchen in der biomedizinischen Forschung gibt, wurde relativ wenig Aufmerksamkeit darauf gerichtet, wie Tiermodelle unsere Annahmen und Erkenntnisse über die menschliche Kognition beeinflussen (Akhtar, 2015). In dieser Arbeit zielen wir darauf ab, die methodischen Lücken und Einschränkungen von Tiermodellen bei der Untersuchung der menschlichen Kognition anzugehen. Dazu haben wir eine rigorose und kritische Analyse zweier Fallstudien durchgeführt, die unterschiedliche Anwendungen von Tiermodellen in der Kognitionsforschung veranschaulichen. Unsere Ergebnisse heben die grundlegenden translationalen Herausforderungen hervor, die ein solcher Ansatz mit sich bringt, und unterstreichen die Bedeutung einer Verlagerung hin zu Methoden, die sowohl empirisch als auch ethisch robust sind, um ein solides und zuverlässiges Verständnis der menschlichen Kognition zu gewährleisten.

1. Introduction

Throughout history, humans and animals have shared a complex and multifaceted relationship. From being a threat to humans as predators, to eventually helping us protect and supplying our societies as we were able to domesticate some of them, animals have always played a crucial role in our development as a species. Additionally, animals have also played a significant role in the cultural and spiritual aspects of human societies. Many cultures consider certain animals as sacred or have developed deep connections with some of them, which consequently has influenced their art, literature, and beliefs. For instance, according to Brian Morris (2006), animals often had a central position in human religions and belief systems. He points out how ancient Egyptians worshiped cats as a symbol of divinity, or that in Hinduism cows are considered “the embodiment of the divine mother”. Other scientists, such as Wendy Doniger (1980), add to this view by arguing that the symbolic significance of animals is not only limited to human beliefs and the development of religions but that animals are also often reflected in our art and literature.

Nowadays, perhaps the most prevalent view on how animals participate in our lives is mainly connected to their companionship, food source, and labor-related activities. However, it is important to emphasize that in the last centuries, animals have also been heavily used in science as analogies for humans. A common example comes from the pharmaceutical industry, where the assumed similarities between different animal species and humans are considered foundational so that newly developed medical drugs can be tested. In scientific language, this trans-species comparison is also known as animal modeling or animal models. In essence, this methodological approach relies on the assumption that animals can serve as reliable representations of human biology, behavior, or cognition and that the insights gained through animal experimentation are applicable and can be translated to humans.

The contribution of animal models to science cannot be simply disregarded. In fact, it is worth noting that nearly 90% of Nobel Prize research in Medicine and Physiology has involved the use of animals in their experimental setups (Andersen & Winter, 2019). For example, the development of insulin therapy for diabetes was achieved with the help of dogs used as models (Banting et al., 1922). Additionally, modeling animals has also been a crucial method when it comes to the development of various vaccines, which ultimately saved countless lives (Kessler et al., 2011). Today, the collective scientific efforts that aim to better understand humans are perhaps more dependent on animal research than ever before, especially when it comes to finding treatments for

the thousands of diseases and viruses that can afflict us. For instance, only in the United States it is estimated that the annual total vertebrate animal usage in biomedical research and teaching was 15 to 25 million per year during 2001–2018 (Kinter et al., 2021). However, despite these extremely high numbers, the progress made in the development of better drugs and treatments seems to be stalled (Hyman, 2012). Some researchers, such as Mak & Evaniew (2016), argue that the lack of progress is due to the complexity of human biology and the limited ability of animal models to accurately replicate our nature.

Naturally, the use of animals in research on such a big scale has significant consequences. For example, public awareness is increasing regarding ethical implications related to animal welfare in scientific research. Moreover, more and more meta-analyses of animal modeling in biomedical research suggest that translating the results to humans is often unreliable and unsuccessful (Greek & Menache, 2013). One such example is reported by Avdeef (2003), where he points out that there are significant differences in the way animals' and humans' drug metabolism work, leading to discrepancies in the extrapolation results and measurements of drug efficacy and safety. In addition, there are also concerns about the ecological validity of animal modeling. More specifically, the experimental conditions are often not representative of our natural environment, which can consequently affect the applicability of the study findings to real-life situations (Ioannidis, 2016).

Within the field of bio-medical research, there is already an ongoing debate that questions whether and to what extent animal models are useful to produce reliable and beneficial human outcomes. In his speech on the topic “Systematic Reviews of Animal Experiments Demonstrate Poor Contributions to Human Healthcare” (2016), prof. Andrew Knight highlights some significantly problematic and worrying statistics with respect to the translational value of such research. Unlike standard reviews that focus on discovering how often results from animal models can be successfully applied to humans, prof. Knight approached this task by looking into systematic literature reviews that are related to the usefulness of animal models in bio-medical research. On that note, it is important to mention that there are notable differences between standard and systematic reviews, which are primarily expressed in the way data is obtained, analyzed, and evaluated. Generally speaking, systematic reviews are considered to provide more reliable evidence, as they examine a large number of experiments and deploy selection criteria that eliminate bias via randomization and other methodological means.

Knight (2016) examined 20 systematic reviews on the contribution of animal models to advancing human clinical outcomes, and an additional 7 systematic reviews on the topic of human toxicity classifications. For instance, one review paper analyzed all research publications utilizing animal models that were published in the top seven leading scientific journals (based on impact factor) and had more than 500 citations. According to this selection criteria, 76 animal modeling studies were identified with a median citation count of 889. One might assume, given how many times those studies were cited and the impact of the journals they were published in, that those animal modeling experiments would be some of the best ones, with a high predictability for positive human outcomes. However, the numbers describe quite a different picture. Only 36.8% of these studies were replicated in human randomized trials, 18.4% were contradicted by randomized trials, and 44.7% never made it to human trials. Ultimately, 10.5% (eight) of all 76 animal studies were eventually approved for use in patients (Hackam & Redelmeier, 2006). Nevertheless, it is crucial to mention that even this does not automatically translate into positive human outcomes, as the adverse reactions to approved interventions are the 4th leading cause of death in US hospitals (Lazarou & Pomeranz, 1998).

Overall, Knight (2016) is one of the many voices in bio-medical science that provide a comprehensive analysis and data that suggests how animal models are more of a burden to scientific advancement in the field, than a method that contributes to it. Another example that the author highlights, is that out of 53 studies that used animal models to investigate cancer, 89% failed replication with other animal populations. Additionally, research of 100 potential drugs that target ALS (Amyotrophic Lateral Sclerosis) were tested in animal models, but only 8 of them made it to the clinical trials. Professor Knight points out that the total estimation of wasted money in unsuccessful animal modeling research accounts for billions.

1.1 Animal Modeling Within the Field of Cognitive Science

The biomedical field is not the only one where animals are used as analogies to humans. Another scientific area that is heavily based on experimental setups with animals is cognitive science (Palmer et al., 2021). However, the approach towards studying cognitive phenomena through experimentation is naturally different than studying the effect of a drug. While within the field of bio-medical research, the focus is often solely on the physiological responses of the subjected individuals, cognitive experimentation often delves into the domains of perception, learning, memory, and the overall exhibited behavior. For instance, one of the most groundbreaking

scientific findings in the 20th century was the discovery of “classical conditioning” by Pavlov, for which he received the Nobel Prize in Physiology in 1904. By studying the digestive systems of dogs, Pavlov demonstrated how learning and behavior can be influenced by environmental stimuli. His discoveries greatly impacted the development of psychology as a scientific field, which later on also reflected on cognitive science. Many scientists such as B. Skinner, who is considered by many as one of the fathers of behaviorism, were heavily influenced by Pavlov’s findings. In fact, Skinner expanded on classical conditioning with the concept of operant conditioning, which describes how behavior can be shaped through reinforcements and/or punishments. Similarly to Pavlov’s approach, animal experimentation and the generalization of results on human behavior played a key role in Skinner’s research practice.

Perhaps the most important question when we consider cognitive animal modeling is – what is the exact cognitive phenomenon that is being investigated? Each cognitive ability, from basic functions such as attention, memory, perception, and learning, to more complex processes such as planning, decision-making, and problem-solving, is usually studied through different methods. In other words, cognitive scientists use a broad range of metrics and approaches to better capture, describe, and study those functions. From a methodological perspective, such a diversity of methods and metrics can create some considerable obstacles when it comes to linking together different cognitive research findings. Additionally, most tasks designed to measure cognitive abilities and performance are intended for humans. As a consequence, when scientists utilize animal models, sometimes they must adapt the already existing human-based tests. Such practices are rarely robustly validated, meaning that the adaptation of the measurement tool is often simply assumed to function in accordance with the intention of the experimenters. As Palmer et al. (2021, page 2) stated:

> “Many standard tasks used to assess cognition in animal models have been designed to measure cognitive constructs similar to human cognitive constructs, including tests of learning, memory, behavioral flexibility, etc. There remains a degree of uncertainty, however, about whether a given test recruits the same cognitive processes in animals and humans, especially when animal and human tests can differ profoundly in their levels of stress (e.g., due to perceived danger), responses (e.g., large-environment navigation), and task motivation (e.g., escape from electrical shock). Clearly, if the constructs that are being measured in animal tasks do not represent the constructs tested in human tasks, then animal-

human translation will be very difficult to achieve (Deslauriers et al., 2018; Driscoll & Barr, 2016; Flores et al., 2018; Insel, 2010; Malkesman et al., 2009; Wallace et al., 2015).”

Investigating cognition through animal modeling often includes training of the subjected individuals to solve puzzles, navigate mazes, or exhibit particular behaviors (MacLean et al., 2014). Given that cognition is by nature a multifaceted phenomenon, the involvement of several fields and the incorporation of different methods is often a necessity, in order to ensure a comprehensive understanding. Moreover, the interdisciplinary research approach is central to cognitive science and often requires the interpretation of several types of data, such as the brain activity of the animals altogether with their consequent actions in a specified task.

From an interdisciplinary perspective, each of the six pillars in cognitive science (psychology, philosophy, linguistics, neuroscience, artificial intelligence, and anthropology) has different methods through which these subjects generate knowledge. While psychologists are often relying on behavioral experimental setups by isolating a variable in the environment and manipulating it, neuroscientists are applying brain scanning and neuroimaging methodologies, to better understand how the brain functions. Those differences are at the same time enriching the field of cognitive science but also causing considerable difficulties. On the one hand, the discoveries within cognitive science can be more robust, plausible, and ultimately reach a greater complexity when several fields and methodologies investigate the same phenomenon. This being said, bridging the gap between those methods proves to be often an extremely challenging task. This is at least certainly the case when it comes to animal models, where animals are experimented on in both of the occasions just discussed. One common issue that occurs both in the bio-medical and cognitive science fields with respect to animal modeling is that the ethical standards for animal experimentation are by default lesser than the ones where human participation is taking place, leading to animals being often put through severe stress, pain, and suffering. For example, animal modeling within the field of neuroscience often apply invasive methods as standard procedures. This could include brain surgeries, in order to install microchips in the animal brain for tracking brain activity, or even intentionally harming/removing brain parts associated with a given cognitive functionality, in order to observe behavioral aspects that are later on translated to humans (Bains et al., 2016).

The wide range of animal species that are used for translational research can also add considerable complexity when trying to understand cognitive phenomena. For instance, primates

are much closer to humans than rodents in the way their emotional behavior is expressed. This makes them a preferred candidate when investigating social cognition due to the assumed extrapolation ease. On the contrary, most of the memory-related modeling experimentation is done with rodents, partly because the “state of art” research on memory today is dominated by neuroscientific methods. It is obvious that the extrapolation strategies from rodents to humans and from primates to humans should be quite different considering the evolutionary, physiological, and psychological closeness when results are translated. Moreover, most cognitive scientists agree that our cognitive domains, whether they are related to memory, language, thinking, consciousness, emotions, etc., are interlinked and have the ability to moderate, influence, and interact with one another. For example, if the emotional states of the organism are actively influencing the perceptual domains, then translating perceptional research from primates to humans should be advantageous compared to rodents, given that humans and primates share much more similarities in terms of emotional display. By studying different cognitive domains independently, scientists’ wish to integrate knowledge into a general cognitive framework faces considerable limitations, and the fact that most results are obtained through different species seems to introduce yet another obstacle in reaching that goal.

1.2 Reliability of Animal Models as a Method

The development of animal models with the goal to better understand human cognition is a challenging and ambitious endeavor. There are many methods and approaches one could undertake, in order to tackle this cumbersome task. However, before going into the details regarding how animal models are, or should be, actually validated, I shall at least partially describe the main premises, assumptions, and expectations one has, when one decides to acquire knowledge through non-human modeling. And indeed, the key word here is “model”.

Firstly, one should assume that there are fundamental similarities across species when it comes to cognitive functions. In other words, there is no way to build a translatable model in a non-human animal, if we believe, or even better, we already know that the given animal has a fundamentally different cognition than us, humans. For instance, let us consider bats and the way they process auditory stimuli. It is well documented in numerous studies that bats rely almost exclusively on their auditory system to navigate through space and build their perception of the surrounding environment. The cognitive mechanism that enables this ability is called “echolocation”. By making specific sounds, bats are able to process the returning vibrations of

sound waves in a way, that allows them to accurately estimate the type of objects around them, as well as their bodily position in relation to these objects. Now, I would not surprise anyone if I say that our human auditory system is substantially different to what I just described. In that sense, a translatable bat model of our auditory system would not be reliable, to say the least. I would even go further and say that a bat model, where the auditory system is a fundamental part of higher-level cognition, would also be doomed to fail any translation attempts. Perhaps, this is one of the main reasons why we don't see bat models dominating translational experimentation. On the other hand, other types of animals such as primates, rodents, birds, cats, and dogs, have been, and still are heavily used for the creation of various cognitive models (i.e. to research social cognition, attention, memory, or learning). The implication of such a trend should be therefore clear – the more often an animal type is chosen for creating a model, the stronger the assumption that humans, and the chosen animal, have fundamentally similar cognitive functionality.

Some of the most important questions that anyone who pursues science should ask oneself are:

- Am I using the right method to investigate the phenomenon of interest?
- How reliable is the method I have selected?
- Is the data that I have acquired relevant for the phenomenon that I intend to investigate and how much this data can actually tell me about that phenomenon?
- What conclusions am I empirically allowed to make, given the methodology I have used and the results I have acquired?
- Is there any bias in my assessment and have I taken all the steps necessary to eliminate it?

When carefully addressing such questions, one assures the quality of the study, how robust and valid would be the acquired knowledge, and how much one can generalize from the specific testing population to the masses. Modeling animals, in that sense, is no different than any other methodology and those questions are equally essential, as they would be if the scientific method was a survey or a questionnaire. The reasons for that are many, but perhaps one of the most important is that more often than I would wish for, using animals to model humans also poses serious ethical considerations.

Most scientists would agree that to pursue a successful assessment of the reliability of any animal model, one must first thoroughly examine not only the methodological foundations, but also their limitations and the assumptions one has and relies on. In order to do so, two main frameworks shall be applied throughout this evaluation, in order to ensure the plausibility of the conducted research – internal and external validity. These two concepts are the cornerstones upon which one can determine the robustness, applicability, and reliability of findings based on animal models. It is important to outline the role of both concepts and how they relate to experimental setups, especially when there are considerable ethical implications of using animals for research purposes. Therefore, I shall highlight the distinct characteristics of the two validity notions, in order to facilitate a better understanding of how experimental research should be validated.

Internal Validity

Internal validity is a concept designed to test whether there is an actual causal effect between the dependent and the independent variable/s in an experiment (Slack & Draugalis, 2001). Through internal validation, one can assess if the observed effects are genuinely due to the manipulation of the independent variable, or a result of the experimenter's biases and/or other confounding factors that haven't been addressed. Establishing a robust and reliable framework that tests for bias, significance of the observable effect, the characteristics of the testing population, and other potential confounds, is crucial for any research that aims to provide accurate and credible findings.

Confounding factors that remain unaddressed can seriously hinder the reliability of any experimental setup. A good example of scientific conclusions that are based on evidence that failed to meet internal validation is described in the paper "Learning Styles: Concepts and Evidence" (Paschler et al., 2008). In a nutshell, the authors discuss whether there is solid scientific evidence for people having different learning styles (i.e. visual style, auditory style, kinesthetic style) and that targeted teaching based on those styles can lead to better learning outcomes for the learner. Although this hypothesis was investigated and confirmed by several studies, the authors suggest that all of them failed to meet the internal validity criteria due to lack of sufficient control for confounding factors. More precisely, they argue that from the whole literature on that phenomenon, very few studies used an experimental methodology capable of testing the validity of learning styles and that there is virtually no evidence for testing the interaction variable, which is crucial in a learner-teacher relationship (Pashler et al., 2008).

Even though addressing confounds is fundamental for the validity of any research, they are not by any means to the only component that can hinder internal validation. Handling confounding factors is just one of the seven pillars upon which internal validation is based (Supino, 2012; Walach et al., 2006). Researchers are expected to not only critically evaluate the possibility of confounds to occur, but also to follow procedures that can strengthen the reliability of their experiment such as randomization, control groups, blinding, standardization, sample sizing, and counterbalancing. A closer look into each of these criteria is necessary, in order to understand why they are important and how they can enhance the reliability of any experimental setup.

When it comes to randomization, it usually means that the experimenter should randomly assign subjects (in our case, animals) to the predefined experimental groups. This is essential because it controls for individual differences that can occur in any population. In that sense, randomization supports the even distribution of such differences across groups. If followed rigorously, this simple procedure also eliminates the potential for selection bias in the experimenter and ensures that the experimental groups are comparable. Additional security in terms of internal validity here can be added by control groups, whose main function is to provide a benchmark versus the experimental group. By having a control group, one can track if the treatment in the experimental group is causing the observed effects, rather than other unrelated factors that were not the focus of the study. Both randomization and control groups should be defined before the experiment begins. However, a third factor is also crucial when considering the empirical preparation, namely, sample size (Rouqette & Falissard, 2011). This might seem like a self-explanatory concept, but there are important factors to consider when addressing the size of the sample. Questions related to the expected scale of the effect that is under investigation, or how broadly one wants to generalize the observed effect, could all be reliably addressed with an adequate sample size. As a general rule, a smaller sample size can still be representative if the observed effect is immense and statistically powerful. On the contrary, if the statistical power of the observed phenomenon is not that strong, then a much bigger sample size is needed, in order to robustly confirm the consistency of that effect over a larger population.

Even if the preparation for the experiment followed the above-mentioned procedures, there are still some important actions that could support the reliability of the outcomes during the time of experimentation. Blinding is one such action that is specifically designed to minimize the experimenter’s bias and participants' expectancy effects (Li et al., 2023). The main execution of blinding consists in the unawareness of the experimenter and of the participant regarding group

assignments. This way, the researcher ensures that data collection and analysis are conducted without initial assumptions and judgments, by limiting the participant's, or the evaluator's expectations and biases. In addition to blinding, counterbalancing the subjects within the experimental groups, or the sequence of experimental procedures through which these subjects go through, is a crucial measure, especially when it comes to repeatable data collection. This is particularly important when there are multiple conditions. The reason is that, for example, fatigue in the participants could cause confounding effects that skew the data sample. By implementing counterbalancing procedures, the experimenter can control for varying the sequence of conditions across subjects, leading to a lower risk of order effects (Greenwald, 1976). Last, but not least, the standardization of the experimental procedures is a key to reliable experimental results. This would mean that the experimenter should be consistent with his/her approach to all participants, and all experimental groups. Whether it is blinding, counterbalancing, or other internal validation measures, all of them should be applied to every participant and experimental group in the same consistent way.

External Validity

From a broad perspective, the method of external validation encompasses the extent to which the findings of a given study can be extrapolated to different settings, populations, species, and/or conditions (Slack & Draugalis, 2001). External validity refers to how well the results of a study apply beyond the specific context in which the research was conducted. For example, this might involve applying findings from a study on laboratory rats to wild-living rats or generalizing results from a small clinical trial to a broader patient population. The difference from the concept of internal validation, which primarily evaluates the robustness of the scientific experimental design, is mainly expressed by the additional application of external validity processes that aim to introduce another layer of assessment from external perspectives. Such processes might involve seeking expert peer review, or the successful replication of the experiment with different populations from several independent research parties.

External validity is sometimes also referred in the scientific literature as generalisability and translatability. However, Pound and Ritskes-Hoitinga (2018) have stated that this is a common misconception, as the two notions are discrete. External validation/Generalization is the *prerequisite* to translatability and only *contributes* to it. In other words, one cannot reliably translate results from animals to humans if the obtained results cannot be generalized.

Another crucial consideration that falls under the external validation measures is the concept of ecological validity. This notion assesses the extent to which the findings of the study can be generalized to real-world settings. As a considerable high number of transgenic and behavioral animal modeling experiments are conducted in a controlled and artificial environment, it is crucial that there is an established measurement which ensures that the obtained results can be applied beyond the conditions of the study (Brewer, 2000). From a logical perspective, it follows that all animal models should be controlled for ecological validity, as the main purpose of the model in the first place is to prove a point about humans. Moreover, since the extrapolation with animal models is made not only to a different population, but to an entirely different species and living environment, the necessity to implement robust external validation measurements is even higher (Sjoberg, 2017). However, Pound and Ritskes-Hoitinga (2018) distinguish two types of issues that occur when considering external validation, namely, the surmountable and insurmountable problems. They argue that while the surmountable problems of external validity could be resolved (i.e. observing animal behavior in their natural environment), the insurmountable problem – species differences – would always undermine generalizability and therefore hinder the success of translational attempts.

Some scientists attempt to address the problems with external validation by introducing genetically modified animals as a potential method that could enhance successful extrapolation attempts. The idea behind such an approach is that by altering the genes of animals to more closely resemble human ones, scientists aim to improve the accuracy of predictions about how humans might respond to certain conditions and treatments that are tested over the modeled animals. However, even with the introduction of genetic modifications, fundamental biological differences between species may still limit the reliability of such extrapolations. For example, Lynch (2006) suggests that a possible solution to the translation problems that emerge when animals are used as a model could be to empirically prove that a gene that is present in both the modeled animal and humans, has a similar functionality. This, however, leads to the so-called “extrapolation circle”, making the animal study most likely completely redundant (Pound and Ritskes-Hoitinga, 2018). To put it simply, in order to determine whether a gene has a similar functionality within humans and the selected animal, this already assumes that the functionality of the gene on a human level is known and understood. Otherwise, how could one be able to judge whether a comparability is possible? Having said that, if the gene and its functions are already known on a human level, then how the animal study would add value and knowledge? On the other hand, if we assume that the

functionality of the investigated gene/s is not known, then another risk occurs, namely, that one cannot be sure what role exactly the investigated gene plays within the different species. Such issues are important to consider when the method of the model includes genetic modifications, as they pose yet another layer of translation challenges.

Another idea that scientists support is that if issues related to internal validation were resolved (rigorously established control groups, measures to prevent bias, blinding, etc.) and experiments with animal models followed a strict experimental conduct, the translation of animal studies to humans would be more successful (Dirnagl & Found, 2014). However, Pound and Ritskes-Hoitinga (2018) find such claims unreliable. They point out that in 1999, the Stroke Treatment Academic Industry Round-Table introduced a series of standards and recommendations intended to improve the quality of animal studies related to stroke. Unfortunately, thirteen years later the translatability of animal models did not experience any improvement (Sutherland et al., 2012). Such evidence suggests that even if internal validation measures are improved, this does not necessarily have an effect over the extrapolation problem. To illustrate the relationship between internal and external validity, Pound and Ritskes-Hoitinga (2018) propose the following graph:

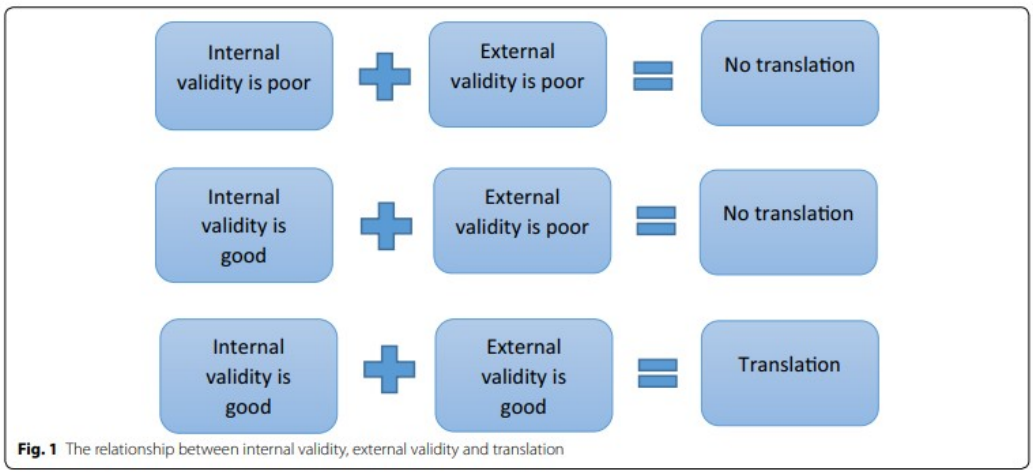


Fig. 1 The relationship between internal validity, external validity and translation

To put it simply, the authors of the unambiguously titled paper *"Is it possible to overcome issues of external validity in preclinical animal research? Why most animal models are bound to fail"* propose that both internal and external validity are necessary, in order to speak about successful extrapolation of findings from animals to humans. Here an obvious question arise – what are the factors and obstacles in obtaining external validation? To ask such questions is important, especially when the evidence suggest that improvements of internal validation measures in animal research do not result in more effective and efficient translation to humans (Sutherland et al., 2012).

To summarize it all, the successful translation from the modeled animals to humans depends critically on both internal and external validity. Internal validity ensures that the experimental design is robust, meaning the results accurately represent the relationship between the studied variables, and that confounding factors are controlled for. In a way, this validation criteria simply ensures that the research conclusions are scientifically sound. External validity, on the other hand, evaluates whether these results can be generalized beyond the specific experimental conditions, including whether findings from animal studies are applicable to humans. Any study that lacks internal validation is empirically flawed and thus a translation cannot occur. However, to pursue a successful translation, researchers should take into account both validation instances, as operating only within the internal validity frameworks would ensure nothing more, but that the generated outcomes are representative for the tested population. There is a considerable amount of evidence suggesting that species difference will always cause issues in translation attempts, regardless of the efforts spent in thoroughly conducting internal validation checks (Pound & Ritskes-Hoitinga, 2018).

1.3 Brief History of Cognitive Animal Models

From a historical perspective, the use of animals for scientific advancement could be probably traced back to ancient times, where dissections helped societies understand how different organs are interconnected, look, and function. However, it wasn't until the early 19th century when animal experimentation began to be used in a more rigorous and empirical manner, especially with the development of modern medicine. One of the earliest examples of animal modeling as a scientific approach that I was able to trace comes from France, where François Magendie (1840) used animal experimentation to advance our understanding of human physiology. To be more specific, the scientist was investigating the function of the blood pressure in our bodies and gained remarkable insights about the role of the lymphatics in the absorption of chemical substances and body poisons, and the separation of sensory and motor nerves in the spinal cord (Stahnisch, 2008). In addition to his epistemic endeavors, Magendie had also practical interests throughout his experimental trials, which can be confirmed by his attempts to develop a blood pressure measuring device.

In parallel with the advancements in medicine, 19th century was also the time when psychology began to establish itself as a stand-alone scientific field. One of the most prominent names associated with this establishment is of Edward Thorndike (1874-1949), who looked into the

similarities between animals and humans beyond the physiological makeup. His main interests were related to, what we would say nowadays, our cognitive functions. The ability to learn and memorize, how associations occur and what is their function, or how we solve problems and make decisions were all problems to which Thorndike dedicated most of his work. Believing in the potential of uncovering more about the mechanisms behind our cognition, he developed numerous experimental setups with animals, in order to understand human intelligence. This is particularly evident in his book *"Animal Intelligence: An Experimental Study of the Associative Processes in Animals"* (1898, page 2) where he writes:

> "The importance to comparative psychology in general of a more scientific account of the association-process in animals is evident. Apart from the desirability of knowing all the facts we can, of whatever sort, there is the especial consideration that these associations and consequent habits have an immediate import for biological science… This, however, is a minor consideration. The main purpose of the study of the animal mind is to learn the development of mental life down through the phylum, to trace in particular the origin of human faculty."

Thorndike was also aware that merely observing animal behavior and drawing conclusions is not a sufficient and internally valid approach to science. Thus, he developed an explicit methodological framework throughout his experimentation, that he continued to follow systematically. He realized that the experimental conditions in which a given behavior is observed need to be consistent, the number of trials should be sufficient, as well as that there should be also a variety of experimental subjects/animals, instead of repeatedly using only one individual. Additionally, he was extremely strict with the potential external influence that the experimenter could have on the animal. Limiting the interaction between the experimenter and the subjected animal to merely moving the animal from its "housing" to the experimental area, allowed several scientists to work on the same research without jeopardizing the study. However, it is important to mention that the personal influence was not the only factor associated with the experimenter that Thorndike tried to control for. He also found for crucial that personal judgments should be eliminated and that the research approach shall be conducted in a way that, citing: "*...the question of whether an animal does or does not form a certain association requires for an answer no higher qualification in the observer than a pair of eyes*" (Thorndike, 1898, page 7).

Another solid evidence for the rigorous methodological approach of Thorndike can also be considered the determination with which he described the experimental setups he used. Below you see a drawing of a box used for experimental setups with cats, purely having illustrative purposes.

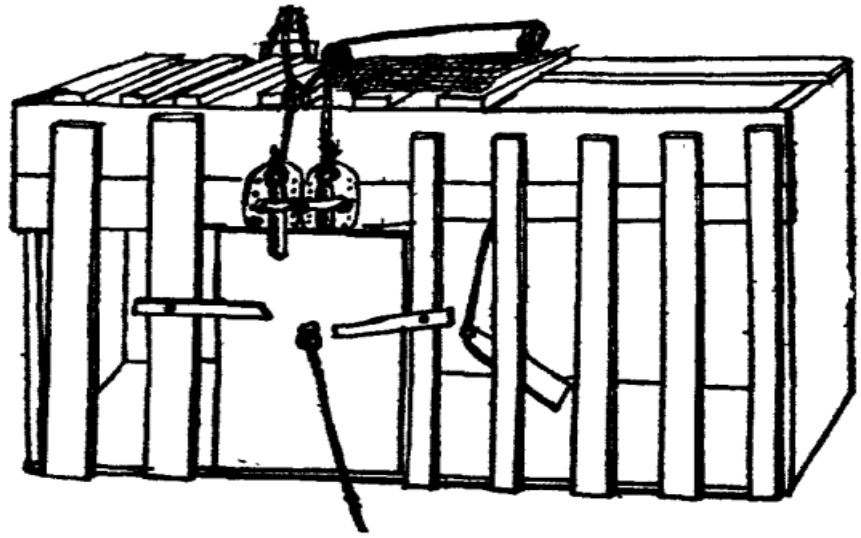


FIG. 1.

Animal Intelligence: An Experimental Study of the Associative Processes in Animals (1898)

The drawings were also accompanied with detailed description of the way the apparatus was built, as well as how exactly should be used. Such pictures underlie the importance of the scientific method in the early developments of psychology as a field. They were also crucial for allowing other scientists to replicate and reproduce the experiments. In today's world of science, such exhaustive and detailed descriptions should be a common practice, especially when one wants to ensure that internal validation procedures are rigorously followed, and also that external validity (through replication) is possible. Considering this, Edward Thorndike played a major role not only in the development of some of the first cognitive animal models, but also in the establishment of robust set of rules which reinforced the scientific method when the objects of investigation were our mental processes.

The 20th century played a pivotal role not only in the development of animal models, but also in how we see science in general today. When it comes to the field of psychology which was later integrated as one of the pillars of cognitive science, one name shines forth as one of the major

contributors, if not the father of the so-called behaviorism – B. F. Skinner. More than a scientific approach, behaviorism was established with the intention to be the only viable and empirical way to study mental processes and cognitive phenomena. This theoretical framework, which evolved into a significant scientific movement throughout the 20th century, claimed that the mind is a "black box" and that the only way to study it, is to observe and measure behavioral output as a result of a controlled environmental stimulation. Inspired by this idea, Skinner revolutionized the field of behaviorism with his invention of the operant conditioning chamber also known as the "Skinner box". In this chamber, different animals such as pigeons and mice learned to associate specific actions, like pecking a lever, with rewards (i.e. food). By experimentally manipulating the conditions, Skinner aimed to investigate fundamental learning processes through reinforcement schedules, which later on paved the foundations of operant conditioning (Skinner, 1953).

Perhaps the most influential concept that sparked the fascination of Skinner with behaviorism comes from one of the key terms in behavioral psychology, namely, classical conditioning. Discovered by Pavlov in the very beginning of 20th century and later on extensively described in his work "Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex" (1927), classical conditioning stands for a type of learning process where an involuntary response occurs as a result of a neutral stimulus due to a formed association. Pavlov's famous experiment involved dogs, where he successfully induced a naturally occurring response of heightened salivation as a result of a neutral stimulus such as a bell ringing. Watson & Rayner (1920) extended the work of Pavlov, providing unparalleled evidence for the manifestation of classical conditioning in humans. For instance, they have demonstrated in a controlled environment that particular neutral sound can be consistently paired with a blink from the participants. In a nutshell, the experiment involved multiple repetitions where participants heard a specific sound followed by a puff of air directed at their eye. After several trials, the subjects were successfully "conditioned" to blink, only as a result of hearing the sound. Such experimental evidence supports the view that some behavioral and cognitive processes could indeed share similarities across species.

Of course, the ultimate goal of Skinner was not simply to provide a behaviorist account for how learning processes and cognition are realized within pigeons and other animal species. He also believed that these basic learning principles could be extrapolated to explain human's learning nature (Skinner, 1974). Inspired by his work, many scientists later on followed the path of behaviorism striving to extend the research scope beyond learning process, and started to develop

new animal models designed to investigate memory, decision-making processes, and even human language. For example, Gardner & Gardner (1969) used chimpanzees that were trained on symbol manipulation tasks to study human language acquisition. Skinner himself also wrote an extensive work on "Verbal Behavior" (1957, page 15) where he claims to provide, citing:

> *"...a way to predict and control verbal behavior by observing and manipulating the physical environment of the speaker."*

This attempt, however, raised a substantial critique from the "cognitive linguistics" camp. One of the greatest scientists in that field, Noam Chomsky, wrote an academic essay in 1959 to address all the shortcomings when it comes to looking at language through the lens of behaviorism. Even more, Chomsky's critique not only defended the idea that language could be to a large extent an innate cognitive property, but also expressed an explicit doubt in the fundamentals of Skinner's behaviorism and operant conditioning. In his work "A Review of B. F. Skinner's Verbal Behavior" (1959, page 26-58), Chomsky writes:

> "We must decide, first of all, whether any physical event to which the organism is capable of reacting is to be called a stimulus on a given occasion, or only one to which the organism in fact reacts; and correspondingly, we must decide whether any part of behavior is to be called a response, or only one connected with stimuli in lawful ways"

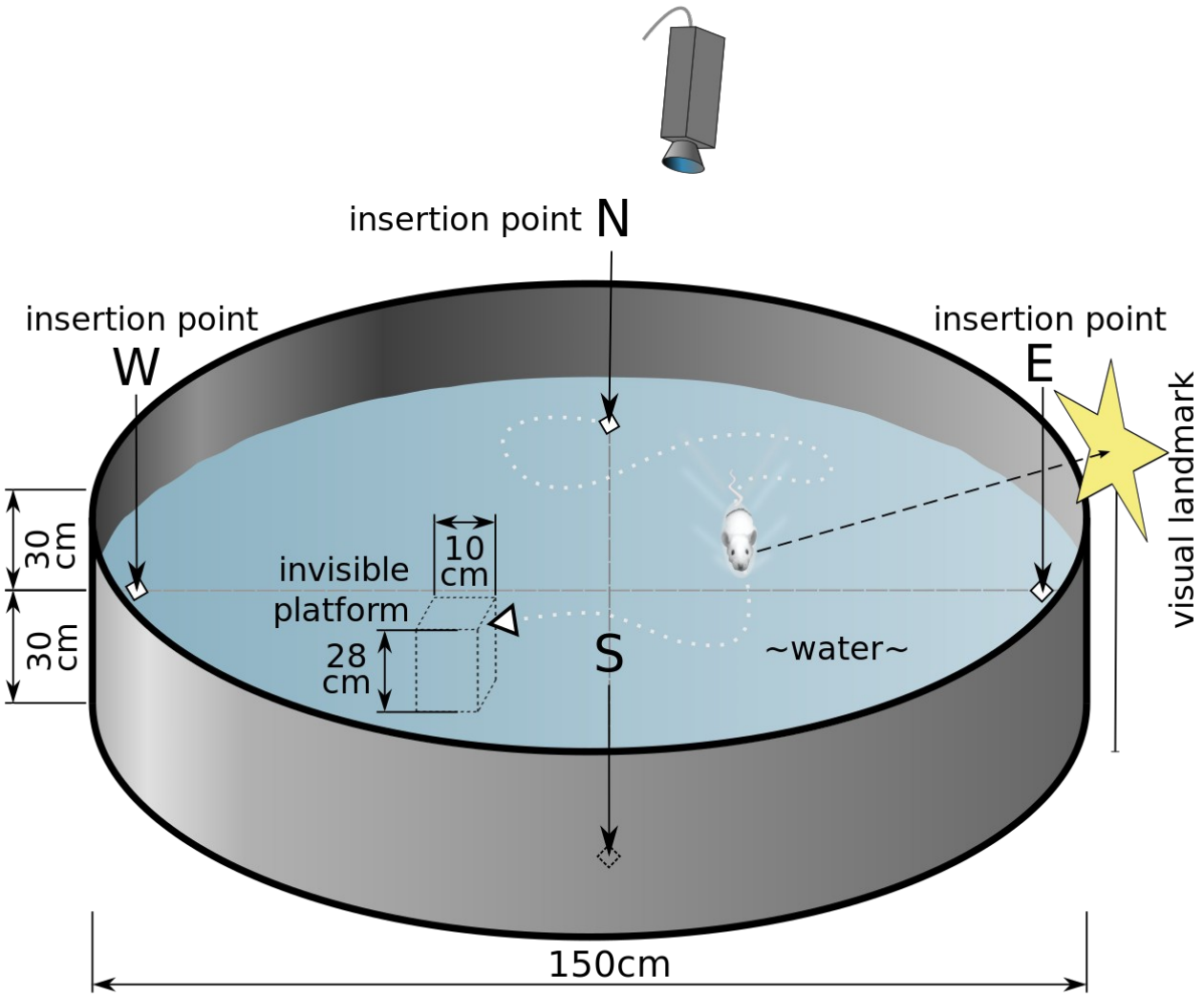
To better illustrate the fundamental gaps with behaviorism, and especially when behavioristic approach is used for explaining language, Chomsky (1959, page 26-58) suggested that the stimulus → response explanatory model cannot provide sound, or sometimes even logical evidence, when one is saying something regarding one's past or imagination. The following citation explores his position:

> "I have often used the words Eisenhower and Moscow, which I presume are proper nouns if anything is, but have never been stimulated by the corresponding objects. How can this fact be made compatible with this (behaviorism) definition? Suppose that I use the name of a friend who is not present. Is this an instance of a proper noun under the control of the friend as stimulus?"

From such examples it is clear that if there are stimuli that are responsible for our verbal behavior, namely – to say Moscow or the name of a friend who is not present, then they could be very much linked to an internal emergence. One of the major points which Chomsky made against Skinner's description of human verbal behavior is the limitations that the behavioristic approach implies by not taking into account any possible internal processes or organization, which may have a causal effect on the way we verbally behave.

While the obvious opposition between behaviorism and cognitivism was emerging, another aspect that is perhaps of utmost importance was long neglected. It wasn't until the end of the 20th century and the beginning of 21st that the problem of **species differences** began to establish itself as a major concern. For example, Gallistel & King (2009) argued that any attempts to extrapolate findings from animals to humans could be misleading due to the vast cognitive and ecological disparities between the modeled animal and us, humans. The authors stated that reducing the conclusions of any animal model merely to behavioral cues is neglecting a series of individual internal factors such as emotions, motivations, past experience, and individual differences. Moreover, Rollin (1995) also pointed out the danger of anthropomorphization when one interprets animal behavior.

I shall come back to our historical journey to the late 20th century, as it is important to mention the emergence of one of the most utilized tools to study spatial memory – the famous Morris Water Maze Test. The invention of Richard G. Morris in the 1980s is associated with the following experimental setup – the so-called water maze is a pool with a circular shape filled with water, where a small submerged platform served as an escape from otherwise potentially endless swimming. Designed for rodents, which were released at various points around the maze's perimeter, their task was to find the platform by relying on visual cues placed on above-water level. In essence, over many and repeating trials, the subjected animals were able to improve their search strategies, indicating the acquisition of spatial knowledge through memory processes.



Source: en.wikipedia.org/wiki/Morris_water_navigation_task

At a first glance, the experimental setup seems simplistic, yet effective when it comes to measuring spatial memory. However, the biggest impact of Morris' experimental setup was not related to investigating spatial memory itself, but in its ability to produce quantifiable data. Variables such as latency to find the platform, what is the length of the path to reach it, and time spent in the previous platform quadrant after its removal were all some of the key metrics, that were used to provide objective empirical information that can be related to the learning mechanisms of the animals. The standardization of the experimental execution and availability of objective and quantifiable data allowed for the test to be reproduced and compared across studies. With that, scientists moved away from mere measurements of behavioral output, and looked deeper in what was believed to be a "black box", namely, the brain. Measuring rodents with lesions in specific brain regions and comparing them to healthy control groups, as well as the manipulation of

neurotransmitters, allowed researchers to reach the neural basis of spatial memory cognition more closely than anyone before (D'Hooge & De Deyn, 2001). This being said, even the introduction of objective and individually independent variables such as latency to find the platform or the time a subject spent on its previous place in the water maze, were still not sufficient to address generalization issues emerging from species differences. Despite its simplicity, the water maze test couldn't answer the question of whether specific behaviors and cognitive strategies were ultimately driven by innate responses to danger, motivations to escape, or bodily motor capabilities. This problem is still a prominent one even today, making the translation of findings a challenging and empirically questionable endeavor.

At the end of the 20th century and the beginning of the 21st, advancements in technologies and neuroscience allowed researchers to tap into new approaches when developing animal models. Electrophysiological testing gained more and more popularity among researchers, where advanced invasive techniques allowed for single neuron activity measurements. The early pioneers Hubel & Wiesel (1968) made a breakthrough with their single-unit recordings of specific neurons in a monkey brain by finding individual neurons that are responsible for the visual processing of lines and edges. In 1981, their research was recognized with the Nobel Prize in Physiology and Medicine and revolutionized the way we understand vision. Their influence was so big, that many scientists afterwards continued to investigate human cognition through electrophysiological animal models, including birds, primates, and rodents.

Di Pellegrino et al. (1992) used single-unit recordings to discover the so-called "mirror neurons". In a nutshell, their findings focused on singular neurons which fire both when the subjected animal (a monkey) executed specific action and also when the animal have observed another individual to perform the same action. These findings sparked a lot of curiosity and opened the door for further research investigations related to the specific functions of mirror neurons. Although the initial study was focused on the motor system and action understanding in monkeys, scientists quickly began to build hypotheses about the potential involvement of mirror neurons in emotional mechanisms such as empathy, as well as behavioral ones such as imitation (Iacoboni, 2009). The extrapolation to humans happened almost seamlessly, as researchers curiously began to explore mirrored brain activity in different tasks involving imitation and empathy with humans. In fact, those efforts continue even today, as more and more publications are dedicated to provide insights into the role of mirror neurons in our human brain. For example, in their paper *"Neural mirroring mechanisms and imitation in human infants"*, Marshall & Meltzoff (2014) investigated

how mirror neurons contribute to imitation in humans infants through EEG. Another instance of scientists' ongoing dedication to uncover the mysteries of mirror neurons is constituted in the work of Bekkali et al. (2020) through a comprehensive meta-analysis. In their paper *"Is the Putative Mirror Neuron System Associated with Empathy? A Systematic Review and Meta-Analysis"*, the researchers examined the nuances of individual neuronal activity associated with empathy, showcasing the continuous interest of researchers towards this neuronal phenomenon.

By having all of the ongoing developments regarding mirror neurons in mind, it is important to mention that the methods used when mirror neurons are studied in humans differ quite substantially from the single-unit recording approach developed by Hubel & Wiesel (1968), which was later on utilized by di Pellegrino et al. (1992) to find their existence. Instead, the methodology used with humans is related to functional magnetic resonance imaging (fMRI), electroencephalograms (EEG), transcanial magnetic stimulation (TMS), or a different combination of the mentioned. In fact, there is not a single study according to my knowledge, where the single-unit recording method was used, in order to prove the existence of mirror neurons in humans. And of course, this is not a surprise, given that a single-unit recording requires a brain surgery, so that the subjects are prepared for experimentation. Additionally, all of the above-mentioned approaches (fMRI, EEG, and TMS) cannot provide evidence, data, or recording, of single neurons. At best, they can only give insights into the activity within whole brain regions, and large neuronal clusters (Logothetis, 2008; Hallett, 2007). The danger of building an entire scientific field on an unproven concept which is based entirely on the assumption that if monkeys have it, humans have it as well, should not be neglected with ease. Ignoring such a huge issue could not only result in conducting a false-positive science, but also in subjecting other animals and humans to experimental conditions that might be unpleasant for the individual and/or require ethical considerations.

In the last two decades, the scientific scope of deploying animal models continued to broaden. While the core cognitive functions such as attention, memory, and learning are still being investigated through modeling animals, researchers have considerably expanded the range of topics and theories that adopt this methodology. This is particularly true with primates, as biologically and genetically they are the closest species to us, humans, which makes them a preferred choice for modeling human cognition. Recent studies that utilize monkeys have already begun to explore complex domains such as decision-making and social behavior. A prime example for that is the paper "Variation in primate decision-making under uncertainty and the roots of human economic

behavior" by De Petrillo & Rosati (2021), where a parallel between non-human primates' risk preferences and human economic behavior is drawn.

In a nutshell, De Petrillo & Rosati (2021, page 3) compared outcomes from studies that examined non-human primates' decisions in the context of risk and uncertainty, and compared the results to research conducted with humans on a similar topics. The findings were then translated to an economical context, in a way that risk preferences were linked to decisions that evaluate potential gains or losses. However, measuring animal decision-making and linking it to human economic behavior and environment is not a straightforward task. As the authors state:

> "Human preferences can be measured in experiments involving choices between different options, often involving monetary rewards, or by looking at a person's 'revealed preferences' by observing their real-life consumer behavior. Most animal studies, in contrast, use a series of experiential foraging decisions about consumable rewards such as food to infer preferences."

An example for a task that is supposed to measure risk preferences and decision-making can be given within the domain of "learned associations", where animals are first trained that a container with specific color or location could contain a constant or varying number of food rewards. For instance, a blue container might have always three food rewards, where a red one could have either one or five rewards with a 50% chance. Then, animals are tested about their preference of container so that conclusions regarding their risk inclinations can be drawn (Kacelnik & Bateson, 1996).

Based on several studies that incorporated such and similar experiments across different primates (including chimpanzees, bonobos, and capuchins), De Petrillo & Rosati (2021) concluded that emotional states, social context and some neurotransmitter systems can modulate risk preferences in primates like in humans. Furthermore, the authors suggested that such findings indicate how some of the cognitive and emotional mechanisms which drive economic behavior in humans are also shared with other primates. However, even though De Petrillo & Rosati found a resemblance between humans and non-human primates in some decision-making mechanisms, it is crucial to acknowledge the contextual and species leap they made.

One of the major problems with such extrapolation efforts lies, not surprisingly, in the external validity of the study. Humans economic decisions and behavior, such as an act of investment, are often driven by long-term planning and projected beneficial outcomes. In contrast, almost all of the experiments conducted with non-human primates involve immediate tangible rewards, such as food. Moreover, paradigms that utilize animal models often require a specific training of the subjected animals so that they can effectively participate in the experiment. This process is fundamentally different compared to the natural settings, both of humans and of the apes, where decisions are required. Even though the goal of the study was to seek for some evolutionary roots of human economic behavior, such a reductionist approach clearly ignores the nature and complexity of economics. To put it differently, reducing humans' economic decision-making processes to apes' choice of a food container is quite the oversimplification, where factors such as future projections, market dynamics, and culture are simply not considered.

1.4 Methodological Approaches in Deploying Cognitive Animal Models

As we understood, the choice for studying human cognition through cognitive animal models has some fundamental implications. First, underlying similarities on cognitive level between the chosen animal and humans are crucial. For example, attempting to build a reliable model on an animal that has considerably different cognition than ours, such as bats, would be a doomed endeavor from a generalization perspective. Second, it is essential that the researchers have an already existing knowledge of the studied phenomenon and how it behaves both in humans and the chosen animal, in order to ensure a translatable outcome. Studies such as Pavlov's classical conditioning with dogs which was later on replicated by Watson & Rayner with humans, exemplify the importance of following rigorous and thorough approach regarding the experimental design and principles, in order to ensure a reliable translational value.

By establishing the fundamental principles of creating an animal model, we can now deep dive into the specific methodologies used to study human cognition through animals. In the following paragraphs I will describe two main perspectives towards building such models, namely, behavioral and transgenic modeling. As already mentioned, although the different methodological approaches have the potential to offer valuable knowledge, they are also accompanied by major ethical and/or translational concerns. Having this in mind, let us will now explore the specifics of different cognitive animal models.

Behavioral Animal Modeling

One of the most common approaches in cognitive animal models is when researchers develop a behavioral framework, intended to observe animals' behavior in a controlled environment and extrapolate the obtained results to humans. Such models are truly interdisciplinary, as they can integrate core principles from several disciplines, including neuroscience, psychology, and comparative biology, resulting in a complex experimental setting that is designed to investigate the phenomenon of interest. According to Tamura et al. (1997), behavioral animal models are living systems used to analyze brain-behavior relationships under controlled conditions, for the purpose of studying phenomena occurring in another species. In that sense, we can understand the animal model as system A, which has intimate corresponding similarities with system B. Based on these similarities, which are often based on evolutionary principles, the reasons behind the behavior of the investigated phenomena in system A should be also corresponding in system B. In addition, it is important to mention that any animal model should provide a simplification of complex phenomena. That is, it cannot be that the experimental model is more sophisticated than the "real world" event, as this would hinder the confirmation and/or falsification of the hypothesis, making the research irrelevant for the advancement of science (Massoud et al., 1998). Lastly, as the main purpose of any model is its generalization value, the findings obtained under experimental conditions should also be possible to extrapolate to real situations, as well as to conditions different from those under which the study was performed (Staay, 2006).

The creation of a behavioral animal model usually starts with a hypothesis. As the name suggests, the main observable variable is the behavioral performance of the subjected animals, which often is interpreted in relation to additional neurological correlates. Considering this, it is important to emphasize that the prediction of any hypothesis should be based on variables that could indeed be measured. To some people, such a statement might come as a redundant, but in reality, many studies that have adopted animal modeling as their main methodology have actually started with an evolutionary-based hypothesis. This often is already an issue from the very beginning, as it makes the hypothesis extremely difficult, if not impossible, to confirm or falsify, as the evolutionary reasons for the emergence of one or another cognitive aspect cannot be directly tracked. Of course, this is not to say that hypothesizing about evolution-based relationships should be a forbidden practice in modeling research, yet it should not be the only justification for experimenting with, and potentially harming animals. At least from an ethical perspective, a more acceptable approach would be to base the development of behavioral animal models on hypotheses,

which focus lies on the investigation and prevention of cognitive dysfunctions, mental illnesses, or abnormal behaviors in humans. In such cases, at least one could claim that the harm done to the animal was justified by the significance of expected positive outcome for humans.

Once a hypothesis is chosen, the next step is to define the specific variables which should be investigated, as well as to select the tools (i.e. standardized tests) that would be implemented, in order to support the accumulation of meaningful data. This, however, is more easily said than done. The reality is that mental disorders, cognitive functions and human behavior, have enormous complexity, making it a real challenge to be modeled and simplified, in a way that non-relevant components can be reliably differentiated from the relevant ones, in order to create a controlled experimental environment. As Gottesman & Gould point out (2003), the description of the psychiatric symptomatology is not easily translated into behaviors that can be defined and operationalized in animal models. Additionally, the wide-range use of diverse diagnostic systems might introduce further complications, when it comes to selecting the correct factors underlying behavioral and/or cognitive dysfunctions (Oliverio et al., 1992).

A good example for a behavioral model comes from rodents that are often used to model psychiatric disorders, such as depression or obsessive-compulsive disorders. One of the most used tools to study depressive-like behavior is the forced swim test (FST) (Yankelevitch-Yahav et al., 2015). In this test, rats or mice are placed in a container with a circular shape filled with water, from where the escape for the animal is impossible. The whole rationale behind the experiment is to observe rodents' behavior in continuum by noting when changes in the exhibited behavior occur. Initially, the animals display a rather active behavior, trying to swim and climb out of the cylindrical container. As there is no platform where they can stay or rest, the subjected animals have no other choice but to swim, otherwise they would drown. The main variable that is being measured during this test is the so-called period of immobility, which is interpreted as a form of desperation of the animal, resulting in minimal movements and floating on the surface, where they simply try to keep their heads above the water (Porsolt et al., 1977). In a way, researchers assume that the state of physical immobility is a result of the mental states of despair and hypoactivity. The model is intended to evaluate different drugs and their antidepressant effect, where a successful measurement would be the recording of a reduced immobility period after treatment.

The method behind the Forced Swim Test is described in four main parts, namely, preparation for the experiment, animal handling prior to testing, the training procedure, and the behavioral coding (Yankelevitch-Yahav et al., 2015). The preparation phase consists of two separate rooms – one designated for the actual testing, and one serving as a housing place before the test execution. A transparent cylindrical glass container measuring 50cm in height and a diameter of 20cm should be used for carrying out the main part of the experiment. Drying cages, heat lamps, and heat pads should also be prepared, in order to avoid animal's hypothermia after the experimental procedure finishes. Additionally, video camera(s) should be installed, so that a clear view of the behavior of the animal can be recorded for later analysis. From an animal handling perspective, the housing should be standardized to 12 hours light/dark cycle. Moreover, in the day of the experiment the animals should have unrestricted access to water and food, except for the duration of the actual test. There is a difference between handling rats and mice, as the procedure describes that rats should be handled for about 2 minutes daily, 5 days before the beginning of the experiment.

When it comes to the actual training procedure, Yankelevitch-Yahav et al. (2015) describe the importance to distinguish between the way the experiment is carried out with mice and rats, as there are some notable differences.

Rats

- The experimentation is divided into two sessions that should be 24 hours apart. The first session is scheduled to be 15 minutes, which is considered as a "pretest" phase, while the actual test session is the second one, which is scheduled to be 5 minutes.
- The cylinders should be filled with tap water with a temperature 23-24 degrees (Celsius). Additionally, the depth of the water should be adjusted in a way, so that the animal cannot touch the bottom of the container with its hind legs.
- For the pre-test phase, each rat should be placed in the filled with water container for 15 min, after which it should be taken out and put in a drying cage. The temperature of the cage should not exceed 37 degrees (Celsius) and the health of the animal should be continuously monitored.
- After 24 hours, the rat should be transported to the test environment and left there for at least 30 minutes before the actual behavioral testing begins.

- The testing period is only 5 minutes, where the behavior of the rat is recorded with a camera. At the end of the testing time, the animal is taken out and put in the drying cage again, which should be with the same conditions, as the ones described in the pretest phase.

Mice

- For mice, there is only one session that is 6 minutes long. There is a division into pretest, which are the first 2 minutes of this duration, and the actual testing, which are the last 4 minutes of the total 6.
- The temperature of the water should be slightly higher, namely, 25 degrees (Celsius). Similar to the rats, the depth of it should be adjusted in a way, that the bottom of the cylinder cannot be reached with the animal's hind legs.
- A camera should be adjusted and turned on through the whole 6 minutes period, where the behavioral analysis begins only after the first two minutes have passed.
- After the experiment, the animal is taken out and put in a drying cage, where the conditions should be the same, as described in the rat experimentation.

Lastly, it is important to note how the behavior of the animals is interpreted, as this is the basis of the whole experimentation. For example, an "immobility" is coded when the mouse/rat is floating through the water with the absence of any movement, where an exception is made only for those movements, that keep the nose of the animal above the water. The duration where the animal "struggles" should also be coded, where the behavior is described as quick and rapid movements, as well as when the forelimbs break through the surface of the water. Swimming behavior is also recorded and identified with movements of the limbs in a "paddling" fashion.

The simplicity of the preparation and execution procedures related to the forced swimming test made it a regular and easy pick as an experimental procedure for numerous studies. In a combination with other methods, such as intentionally induced brain damage/lesions or genetically modified animals, researchers have coded the so-called "despair" of countless of rodents, aiming at uncovering the neural mechanisms of depression and other complex mental disorders. Yet, even if the initial design looks appealing, there are fundamental problems that hinder the translational value of FST. First, there are numerous of factors that can influence the so-called "immobility" behavior. As it is the main variable in the experiment, as well as the primary rationale behind connecting the state of immobility to depressive-like symptoms in humans, one would expect that the interpretation of this behavior should be bulletproof. Quite the contrary, this solely behavioristic approach

introduces the potential for many debates. For instance, Fedotova et al. (2005) have suggested that the immobility might actually reflect a coping strategy based on a learning process, indicating that the best behavior for a long-term survival might be the preservation of energy. Additionally, the individual behavior of the subjected animal depends on numerous factors such as swimming skills or pretesting energy levels. Yet, from a scientific perspective, it is important to zoom out from all the possible arguments in both directions for a second and see the bigger picture. One thing is absolutely clear – there is absolutely no way, tool, or measure, that can empirically and robustly prove whether the subjected animal/s experience a state of despair. That is, the mental state of the animals is simply not reachable. And it is either extremely bold, or extremely ignorant, to test a hypothesis through another hypothesis. In other words, one cannot reliably uncover whether a certain drug has an effect on depressive-like symptoms, where there is no clarity, and most importantly – no way to prove that the exhibited immobile behavior is in fact a depressive symptom.

To summarize it all, behavioral animal modeling is a methodological approach used in cognitive animal research, where animal behavior is observed in controlled settings, with the aim to gain insights into human functionality and cognition. Often these models integrate disciplines such as neuroscience (i.e. when one intentionally damages the brain of the animal to observe the resulting behavior), comparative biology, and psychology. When hypotheses are entirely evolutionary-based, the clarity and reliability of the models could become doubtful and reduce the scientific relevance of the study. Moreover, multiple challenges and issues in terms of behavior interpretation can emerge, due to the fact that there are numerous and diverse factors that can influence how the subjected animal acts.

An example for a behavioral modeling was put forward with the Forced Swimming Test (FST). Despite its appeal, the reliance of behaviors such as immobility as key metrics, which is merely interpreted as an exhibited state of despair, seriously hinders the translational value such models could provide. The individual differences of the animals seem to be omitted, while it is also not clear whether the immobility is a coping strategy, a behavior that emerges as a result of a learning mechanism, or else. Uncertainties such as those pose not only validity and methodological concerns, but also raise serious ethical questions, especially when we consider the suffering of the animals along with the low translational outcomes.

Genetic Animal Models

A different approach to study human cognition is when the modeled animals undergo targeted genetic modifications before the actual experimentation begins. In a nutshell, those models represent an approach where animal's genes are manipulated, often by either modifying specific genes, introducing new ones, or removing such. The main goal of these modifications is to mimic genetic conditions in humans, so that the effect of genes over cognition can be investigated. One of the most popular research domains, where genetic animal models are utilized, is the field of neuropsychiatric disorders such as schizophrenia, Alzheimer's disease, and obsessive-compulsive disorders (Nestler & Hyman, 2010).

The procedure of developing a genetic animal model often starts with the selection of an appropriate species that is believed to share similar genetic composition with humans. Rodents, such as mice and rats, are particularly preferred due to their short reproductive cycles, well-mapped genomes, and the possibility to apply advanced genetic engineering techniques (Rosenthal & Brown, 2007). After the preferred species is selected, then researchers decide what genetic procedure to apply, based on their research goal. For example, the method of transgenesis represents an introduction of new specific gene/s into the genome of the animals, where the so-called gene knockout method deletes targeted genes from the organism (Capecchi, 2005). Given that the desired modification was successfully integrated within the animal's organism, the next step is to breed these genetically altered individuals so that the modified genes are expressed in the next generations. This way, researchers try to control that the changes they observe in the behavior and/or brain function of the animal are because of the genetic alternation and not influenced by unaccounted factors. In terms of the way how scientists register a behavioral change or anomalies, it usually involves the comparison between the genetically modified animals and their wild-type relatives which serve the purpose of a control group.

When it comes to the testing environments in which genetic animal models are deployed, there are several widely spread variations that researchers often use to investigate the potential effects that specific genes might have on behavior and/or brain function. For instance, advanced imaging techniques and the implantation of electrodes into the brain of the animals are a common way to study specific neural structures and circuits which are associated with a particular cognitive function. Additionally, almost any behavioral test can be in theory applied to genetically modified animals, as the changes in the observed behavior are interpreted as an effect that occurred because

of the caused genetic modification. The Forced Swim Test (FST) that I have discussed earlier is a good example for such a testing, where transgenic mice and rats were used to test the effect of different genes that were linked to depressive-like behaviors. For example, Holmes et al. (2003) investigated the specific role of serotonin transporters in the context of FST depressive-like symptoms by deploying mice with modified expression of the 5-HTT gene.

Despite its appeal, genetic animal models face yet again the same considerable limitations, especially when it comes to the extrapolation attempts. For example, human brains and cognition are far more complex compared to other animals that are usually used for transgenic modeling (i.e. rodents) and are influenced by variety of factors that simply cannot be replicated in a laboratory setting. Ecologically valid aspects such as culture, environment, a unique and expanded neocortex, and cognitive development stages, are all playing a vital role in forming our cognition and behavior. However, all of these are disregarded in experiments such as the Forced Swim Test, leading to tons of scientific effort that has little to no translational value. In 2012, Steven Hyman published a strong critique regarding the state of current translational experiments. In his article called "Revolution Stalled" (page 2), the author states:

> "In the widely used forced-swim and tail-suspension tests, for example, a compound is considered to be an antidepressant when, compared with a control, it causes a rat or mouse to continue swimming or actively struggling longer. Forced-swim and tail-suspension tests do not even model the therapeutic action of antidepressants, because in these rodent screens, a single dose of antidepressant is active, whereas in depressed patients, antidepressant drugs require weeks of administration to exert a therapeutic effect…A PubMed search using the terms "serotonin transporter polymorphism depression" yielded 715 papers, and a less-constrained search on the "serotonin transporter polymorphism" yielded 1943 papers. However, recent mega-analysis of genome-wide association studies in human subjects suffering from depression failed to confirm an association with serotonin transporter polymorphisms."

Although the numbers that indicate unsuccessful replication and/or extrapolation within the field of translational medicine are shocking, an exceptional amount of capital, effort, and time is still designated to research that tries to model human cognition in animals. This outlines the need for a better and clearer understanding of the issues with cognitive animal models, even if it is useful to stay on the surface for outlining the problem on a higher level. To better illustrate the limitations

and complexities discussed thus far, a thorough exploration of specific case studies could provide us with more insights and transparency regarding the extrapolation problems that animal modeling research faces. Therefore, I will dedicate my next chapter to two case studies that deploy respectively genetic animal models, as well as a behavioral model as a way to better understand human cognition.

2. First Case Study – Genetic Mice Models of Obsessive-Compulsive Disorders

Paper Title: "*Neurobiology of obsessive–compulsive disorder: insights into neural circuitry dysfunction through mouse genetics***"**

Authors: Jonathan T. Ting and Guoping Feng

Year of Publishing: May 24, 2011

Journal: Current Opinion in Neurobiology (Volume 21, Issue 6, Pages 823-946, Impact Factor: 4.8)

Mental processes, behavior, and the role of the environment in shaping them have always been central topics within the field of cognitive science. Naturally, abnormalities and dysfunctions often provide a useful context for studying the secrets of the mind. By examining deviations, one could reach insights related to a variety of cognitive mechanisms and underpin certain functionalities. Neuroscience alone can offer a humongous amount of knowledge that was acquired precisely by investigating abnormal neural activity or brain dysfunctions. Of course, it goes without saying that the satisfaction of scientific curiosity is not the only reason why mental and behavioral anomalies are studied. Even bigger justification is that they can often lead to suffering, pain, and overall reduced quality of life for those who experience them. This at least is certainly the case when we speak about obsessive-compulsive disorders (Macy A. et al., 2013). People suffering from this condition can often feel trapped and powerless to oppose their own intrusive thoughts, which can lead to compulsive behaviors and the experience of high amounts of anxiety. So far, we have mixed results when it comes to treating this psychological condition and the need to underpin the causes of its symptomatic only grows.

In this section, I will delve deep into a scientific paper that proposes a way to rapidly advance our understanding of OCD. More specifically, the authors suggest that "*the ability to perform invasive techniques in genetic animal models*" is crucial for deepening our knowledge and comprehending the complexity of this mental illness (J. Ting and G. Feng, 2011). The upcoming paragraphs will be therefore designated to thoroughly present the paper in terms of the arguments that are put forth, the underlying assumptions associated with them, and a detailed exposition related to the results, discussion, and conclusions the authors made. However, since the main realm of their investigation is connected to better understanding *"the precise causal factors"* of obsessive-

compulsive disorders (J. Ting and G. Feng, 2011), it is also imperative that we first describe OCD as such, and what cognitive science can reveal regarding this psychological condition.

2.1 Obsessive-Compulsive Disorder – Short Overview

Obsessive-Compulsive disorder (OCD) is a severe psychiatric condition that significantly impairs the day-to-day lives of those who experience it (Macy A. et al, 2013). It is characterized by two main components – obsessions, which are represented by intrusive thoughts over which the individual has little control, and compulsions, which manifest as behaviors exhibited as a form of coping mechanism (T. W. Robbins et al., 2019). The types of intrusive thoughts and the associated with them compulsive behaviors can also differ considerably. People can have different levels of control over the obsessions, as well as a wide range of compulsive behaviors that are used to mitigate the distress. Some of the common examples when it comes to obsessions include the fear of contamination, constant need of checking various things, or increased anxiety related to parting oneself with items of little or no value. The lack of control over such fears and thoughts can then lead to excessive and repetitive behaviors such as handwashing, avoidance of touching public objects and surfaces, constant checking of locks, doors, and technology, and living in cluttered and unsanitary conditions due to accumulation of useless items. It is important to note that those behaviors and thoughts are by no means exhaustive and can be highly complex.

Cognitive-behavioral models provide a useful and descriptive perspective on some of the foundations related to obsessive-compulsive disorder. One such is proposed by Salkovskis (1999), where he emphasizes the importance of ill-based beliefs and appraisals in driving obsessional thoughts and compulsive behaviors. According to his model, people who have OCD often hold irrational beliefs, amplifying the importance of their intrusive thoughts, as well as the danger of not engaging in the compulsive ritual. A proper example of this could be an individual with contamination-based thoughts who believes that touching the fork in a restaurant will inevitably lead to severe illness. As a result, the person could feel almost as if there is no other choice but to immediately perform a compulsive handwashing, in order to cope with the anxiety and stress from the perceived danger. Salkovskis argues that such unstable and abnormal beliefs reinforce the cycle of obsessions and compulsions and enhance the reality of the perceived threat.

Understanding the cognitive-behavioral model of OCD can also provide valuable insights related to the maintenance of the disorder. According to Salkovskis (1999), those who have OCD

often engage in diverse cognitive processes that contribute for the perpetuation of their symptoms. For example, he puts emphasis on the phenomenon of thought-action fusion, where individuals hold the belief that having a distinct intrusive thought is morally equivalent to carrying out the associated action. As a result, people with OCD might have heightened feeling of guilt if they do not perform the compulsive ritual, as well as a strong sense of responsibility. In addition, Slakovskis also points out that individuals who suffer from this neuropsychological condition often exhibit attention biases towards threat-related stimuli, leading to increased alertness and anticipation of distressing thoughts. Such cognitive processes ultimately reinforce the obsessional thinking and the need for compulsive behaviors as a way of their neutralization.

In an effort to understand the underlying causes of OCD, psychiatrists and scientists have classified several subtypes of this mental condition based on the variability of thoughts and compulsions one has. The primary OCD subtypes are therefore contamination OCD, checking, hoarding, ordering and arranging, and relationship-related OCD (Abramowitz and Jacoby, 2015). As the subtypes can suggest, the main cognitive processes that drive obsessions and compulsions also differ. For example, the hoarding subtype can include difficulties in making decisions related to parting with specific possessions while excessive ordering and arranging can be related to extreme attention to detail. Additionally, research on checking OCD shows that it is associated with impairments in working memory capacity, specifically in visuospatial memory (B. Harkin and K. Kessler, 2011). Those differences could be one of the main reasons why a unified and standardized neuroscience-based approach in treating OCD has not been found to date. More specific, the diversity in the cognitive domains behind OCD symptomatic' is creating further complexity, given that the different cognitive functions can be associated with various brain regions and neural circuitry. Therefore, it is crucial for a qualified professional to carefully assess the symptomatic related to whether an individual has an obsessive-compulsive disorder and if yes – what type it is. This rigorous evaluation is necessary to avoid misinterpreting symptoms due to their variability and uniqueness, but also to determine the most suitable treatment or combination of treatments for a specific person. Currently, the medical care of patients who are diagnosed with this mental condition consists of a variety of approaches, including phenomenological treatment, various psychotherapeutic techniques, different types of medications, and even surgical procedures (Franklin et al., 2011).

By now, it should have become evident that obsessive-compulsive disorders are posing several challenges that are quite complex. Identifying this mental condition, categorizing it, trying

to figure out what would be the best treatment, and ultimately – increasing the quality of life of a specific individual who lives with OCD, are all aspects of a multi-faceted problem. One way in which such a sophisticated issue could be approached is by trying to understand the root causes of OCD. For example, it is believed that such an approach holds the potential to enhance the development of better treatments. In recent years, a preferred method to study this psychological condition is by developing animal models. Thus, neuroscience became the field where most of the modern OCD-related research is conducted.

With the current development of technology, neuroscience can employ a wide range of methods to bridge the knowledge gap related to the brain-behavior relationship. However, the methods within the field can be quite distinct and their right selection when investigating OCD could be of key importance. For example, brain imaging methods such as fMRI and EEG could visualize the brain activity while a subject is performing various cognitive tasks. These two methods are also non-invasive, as there is no physical penetration of the body to accumulate data, nor any manipulation of neural structures. It is important to mention that non-invasive methods are also much safer for the individual than invasive ones. This being said, itentional brain damage, genetic modification, and microelectrode installations are common ways to obtain cognitive data. For instance, Jonathan T. Ting and Guoping Feng (2013) believe that precisely such invasive approaches are crucial to advance our understanding of the underpinning causes of obsessive-compulsive disorders. Yet, a crucial question that still remains is – How reliable and translatable are animal models of OCD?

In summary, obsessive-compulsive disorder is a serious mental health condition that negatively affects the quality of life of the people who live with it (Macy et al., 2013). It is characterized by intrusive thoughts (obsessions) which result in repetitive behaviors as a form of coping mechanism (compulsions). There is a variety of different obsessions and compulsions, where the most commonly described examples include checking rituals, contamination anxieties, and hoarding tendencies (Robbins et al., 2019). Models such as the cognitive-behavioral one proposed by Salkovskis in 1999, provide a useful framework to study and understand the mechanisms through which this mental condition unfolds. Salkovskis has put an emphasis on the central role that irrational beliefs have, as well as the thought-action fusion, where a person feels morally responsible for acting upon their obsessional thinking. In addition, other researchers have identified tendencies related to the specific obsessions and compulsions one has, resulting in the identification of several obsessive-compulsive disorder subtypes (Abramowitz and Jacoby, 2015).

Harkin & Kessler (2011) suggested that this heterogeneity might be precisely the reason why an unified approach to treating OCD has not yet been established. As the main objective of studying OCD is to help and effectively treat people who suffer from this condition, neuroscience seems to be the preferred and modern approach to accumulate knowledge. However, the implementation of brain studying techniques comes with several issues, especially when we consider that a large amount of the studies are carried out with animals. From one side, invasive methods are raising significant ethical concerns related to the well-being of the animals. On the other side, even when their well-being is sacrificed with the hopes to achieve new breakthroughs, the past history of translatable animal models has shown to be extremely unreliable.

2.2 Cognitive Science and its Relevance in Obsessive-Compulsive Disorders

Cognitive Science provides an unique environment where obsessive-compulsive disorders can be investigated, by having the ability to put emphasis on the intricate connection between cognition, behavior, genetics, and their underlying neural mechanisms. As mentioned above, OCD is characterized by the constant presence of intrusive thoughts (obsessions) and repetitive behaviors that can look very similar to mental rituals (compulsions) whose primary aim is to alleviate stress and also serve as a coping mechanism. Both obsessions and compulsions are inherently expressed through cognitive processes such as attention, perception, memory, and decision-making, all of which are central topics within the field of Cognitive Science. Additionally, the methods employed in Cognitive Science include behavioral experiments, neuroimaging techniques (i.e. fMRI), as well as the development of computational and animal models, all of which can be used to better understand the phenomena of interest. As the science of the mind is also interdisciplinary by its nature, researchers who are familiar with the theoretical frameworks and methodological techniques in this field could have a unique holistic perspective, allowing them to explore OCD beyond mere symptomatology. For instance, cognitive-behavioral therapy (CBT), which is a crucial part of the treatment offered to those who suffer from this mental condition, encompasses the multidisciplinary principles of Cognitive Science by targeting abnormal cognitive processes through restructuring and exposure-response prevention techniques (Abramowitz, Deacon & Whiteside, 2018). Examples such as this one point out the potential of Cognitive Science not only to contribute uncovering the cognitive underpinnings of mental disorders such as OCD, but also to support an informed and theory-grounded development of therapeutic interventions.

One of the most prominent frameworks within the field of Cognitive Science, the Enactivism theory, could offer an unique perspective through which obsessive-compulsive disorders can be understood. In a nutshell, the enactive theory posits that cognition is not just a mere product of internal mental representations, but instead is grounded in the dynamic interaction between the organism and its environment (Varela, Thompson, and Rosch, 1991). Additionally, Enactivism also pinpoints the importance of our action-perception loop as a crucial mechanism related to the development and realization of cognitive processes. Such a theoretical framework could prove to be very useful when one tries to conceptualize disorders such as OCD. For example, if we are about to describe this mental disorder from an enactive perspective, one could say that OCD is characterized by a distortion between perceptual mechanisms and the corresponding actions. People who have OCD often experience a state, where there is a disconnection between their thoughts, environment, and the actions which help them regaining the power to steer the feelings of threat (i.e. the door is not locked) and reduce anxiety. In a way, compulsive behaviors could be seen as disrupted action-perception loops, where adverse mental and behavioral patterns are enacted within the environment.

2.3 Case Study Overview

Jonathan T. Ting and Guoping Feng dedicated their paper *"Neurobiology of obsessive–compulsive disorder: insights into neural circuitry dysfunction through mouse genetics"* to the fact that we still have little understanding when it comes to the exact causes of obsessive-compulsive disorder. The authors point out already in the very beginning of their work that despite the continuous efforts and years of research dedicated to the cortico-stratial-thalamo-cortical (CSTC) circuitry in the brain, which is suggested to be tightly linked with compulsive and repetitive behaviors, the precise causal factors of OCD remain unknown.

It is estimated that around 1-3% of the general population is suffering from OCD which makes it one of the most prevalent neuropsychiatric disorders (M. Weissman et al., 1994, M. Karno et al., 1988). There are several treatment approaches developed, including cognitive-behavioral therapy, exposure therapy, prescription of specific drugs, and neurosurgery (M. Jenike, 2004). However, Ting and Feng highlight that a significant percentage of the people who have OCD remain unaffected by the currently existing treatments. Moreover, invasive interventions such as deep brain stimulation (DBS), limbic leucotomy, anterior capsulotomy, or subcaudate tractotomy have produced controversial results. Trials of capsulotomy, for instance, have shown that only around 45% of the subjected patients have experienced reduction in the severity of their symptoms

(M. Jenike, 2004). Furthermore, adverse effects such as seizure, weight gain, transient headache, and in rare cases cognitive impairments and personality changes, were also evident after the surgery. Such facts are a clear signal that a better understanding of the causes related to OCD is needed. Considering this, the authors propose that genetic animal models and the ability to perform invasive techniques on them will be crucial to progress towards pinpointing the exact casual factors of this mental condition. In fact, both researchers suggested that the lack of progress in understanding OCD might be prescribed to the somewhat limiting ethical frameworks that currently prevent an easier realization of experiments that include animals.

To pursue an understanding of obsessive-compulsive disorder through invasive genetic techniques, J. T. Ting and G. Feng introduce a theoretical framework that explores the neural circuitry dysfunctions which are associated with this psychiatric illness. More specific, research done with different brain imaging technologies have shown that abnormal activity patterns in the GSTC circuit loop correlates positively with OCD-like behavioral symptoms (R. Albin et al., 1989). Additionally, developments in studies that address the basal ganglia circuitry functions suggest that partially closed loops through the CSTC pathway are sub-serving discrete motor and cognitive functions (G. Alexander et al., 1986). Those findings have led to the hypothesis that hyperactivity of the orbitofrontal-subcortical loops caused by a disruption in the activity of the opposing ganglia pathways underpins the emergence of OCD symptoms (R. Albin et al., 1989). Interestingly, the authors state that this theoretical framework is an “oversimplification” of the exact neural regions and activity that represent obsessive-compulsive disorder, yet remains a useful paradigm to study hypotheses that are linked to circuitry dysfunction in OCD. Additionally, Ting and Feng admit that the above-mentioned hypothesis has not been directly supported, nor refuted, by experimental data. The main suggested reasons for that are ethical concerns related to invasive trials with humans and the lack of “well-established alternative animal models”. Therefore, the authors have turned their attention to recent genetic mouse models in an attempt to provide a comprehensive review of the insights they could offer.

Sapap3 null mice – J. M. Welch, G. Feng et al., 2007

The first promising study discussed in the paper, according to the authors, is dedicated to genetically modified mice which have undergone a genetic deletion of Sapap3 protein, also known as DLGAP3 (Jeffrey M. Welch, G. Feng et al., 2007). Since Sapap3 is highly expressed in the striatum, a subcortical part of the brain involved in various functions and part of the CSTC circuitry,

the experiment aims to investigate the involvement of this gene in OCD-like behaviors. To produce Sapap3 null-mice, the researchers performed homologous recombination in mouse embryonic stem cells (Jeffrey M. Welch, G. Feng et al., 2007). In a nutshell, this is a highly complex procedure where modified embryo-cells are injected into a fertile mouse after which the surrogate mouse mother carry the modified embryos to term, resulting in a genetically modified offspring (C. Tong et al., 2011). It is important to mention that such procedure would remove the Sapap3 gene not only in the striatum, where it is highly expressed, but also in the whole organism.

One week after Sapap3 null-mice were born, they were anaesthetized and placed in a mouse head holder, in order to receive an injection directly into the striatum with GFP lentivirus through 2 sites at 8 locations per hemisphere. This is performed, as the green flourescent protein (GFP) allows scientists to track and visualize specific cells or tissues under flourescent microscope. Additionally, all mice that participated in the experiment received acute sagittal brain slices (approximately 300 micrometers thick), so that a field recording electrode can be placed in the dorso-lateral striatum, and a monopolar stimulation electrode in the corpus callosum (Jeffrey M. Welch, G. Feng et al., 2007).

Following the genetic modification, the performed brain slices to place electrodes, and the received direct injections into the striatum, Sapap3 null-mice were ready and prepared for behavioral analysis. At the age of 4-6 months, the modified animals developed lesions on their head, fur loss around the snout and the neck, and hairless skin under the eyes (Jeffrey M. Welch, G. Feng et al., 2007,).

(Symbol +/- refers to a control mouse, while symbol -/- refers to a genetically modified mouse. This picture is taken directly from the scientific publication that describes the Sapap3 Experiment. For more information, please see Jeffrey M. Welch, G. Feng et al., 2007, "Cortico-striatal synaptic defects and OCD like behaviors in Sapap3-mutant mice".)

Control groups were established where the genetically modified mice were housed alone or with cage mates. No significant difference in their symptomatic was recorded. The authors reported that the cause of the lesions, and the fur loss, was excessive self-grooming exhibited by the mice – a behavior linked to obsessive-compulsive disorder, more specifically to compulsive acts. Moreover, mice that already had developed lesions exhibited a similar excessive degree of self-grooming compared to mice that were yet to develop lesions, signaling that level of compulsive behavior is injurious. It was also reported that increased grooming was evident even throughout the hours when mice usually sleep. All behavioral assessments were made possible through 24h videotaped recordings that were later on blindly analyzed.

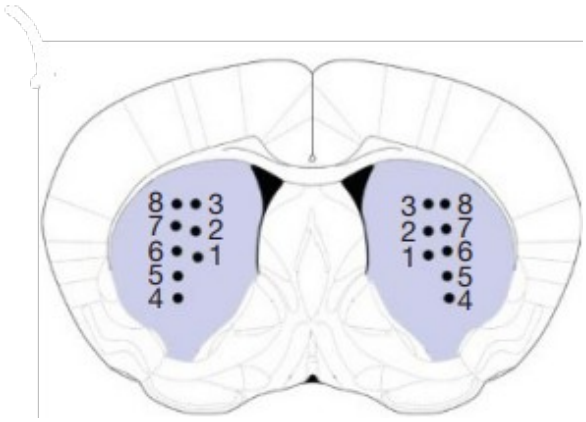
The next symptom associated with obsessive-compulsive disorder that was behaviorally analyzed by the researchers is anxiety. For testing whether the genetically modified animals experienced increased levels of anxiety, standard mice-behavior tests were implemented such as open field, dark-light emergence, and elevated zero maze tests. In the first task (open field), knockout mice have shown to spend less time in exploring the center of the field and more time around the walls, which was interpreted by the researchers as the animals experiencing higher levels of anxiety. When it comes to the dark-light emergence test, Sapap3 null-mice took longer time to pass from the dark chamber to the lit one, while also spending less time in there (an environment, that according to the experimental design is considered stressful), suggesting anxiety reasons. Lastly, the results obtained in the elevated zero maze test have shown that KO mice have spent significantly less time in discovering the open areas compared to wild-type controls, and again took more time to cross into open areas in general.

Evaluation of whether drugs that are currently used to treat OCD would be effective in the symptomatic expressed by the mice was done by the experimenters. More specifically, serotonin reuptake inhibitors (fluoxetine) were used and after 6 days of treatment, anxiety symptoms were reduced. The authors point out that currently such drugs are effective in approximately 50% of OCD patients and the mechanisms of action are not well understood (S. E. Swedo, 2004).

Electrical recordings were also performed to measure how Sapap3 gene absence affects different parts of the brain. To perform the assessment, the authors used a substance called NBQX to block AMPAR synapse receptors, and added another substance called glycine. AMPAR receptors are essential for fast communication between neurons in the brain, while NBQX is a chemical that block AMPAR functions (Petralia & Wenthold, 1992). By blocking those receptors, experimenters

aimed to observe NMDAR functionality – another type of receptor in the brain involved in nerve cells communication. All recordings were done in the absence of magnesium. The obtained results suggest that Sapap3 gene is specifically important for the functionality of the striatum, a part of the brain where this gene is highly expressed. In contrast, the neurotransmission in the hippocampus, another area where Sapap1, 3 and 4 are highly expressed was found to cause defects in synaptic communication. Based on the electrophysiological recordings, the authors conclude that Sapap3 gene has different role in distinct parts of the brain, and the lack of observed deficits in the hippocampus could be due to compensatory effects of other SAPAP members, or redundancy.

Lastly, a lentiviral vector was used as a primary method for gene delivery into the brains of Sapap3-mutant mice. Microinjections were employed to ensure precise delivery of these vectors into the striatum. In an attempt to minimize tissue injury, the experimenters administered one anterior and one posterior injection per hemisphere, reaching a total of eight locations. The lentiviral vectors were designed to transport two types of genes: one for green fluorescent protein (GFP) and another that combined GFP with SAPAP3. Electrophysiological recordings suggested that the GFP-Sapap3 protein bounded successfully with the synapses in the striatum.



The figure shows the locations reached with microinjections per hemisphere. For more information, please see Jeffrey M. Welch, G. Feng et al., 2007, "Cortico-striatal synaptic defects and OCDlike behaviours in Sapap3-mutant mice".)

Four to six months after the injections, all eight mice in the control group (genetically modified mice introduced to GFP) have not expressed a change in their behavior, leading to skin lesions (six of which were severe cases, one was medium, and one had a small open wound). When it comes to the experimental group – three out of eight of the mice had normal skin conditions.

However, the rest still exhibited severe grooming and had open wounds despite of the Sapap3 introduction, where two of them had less severe wounds, although obsessive grooming was evident.

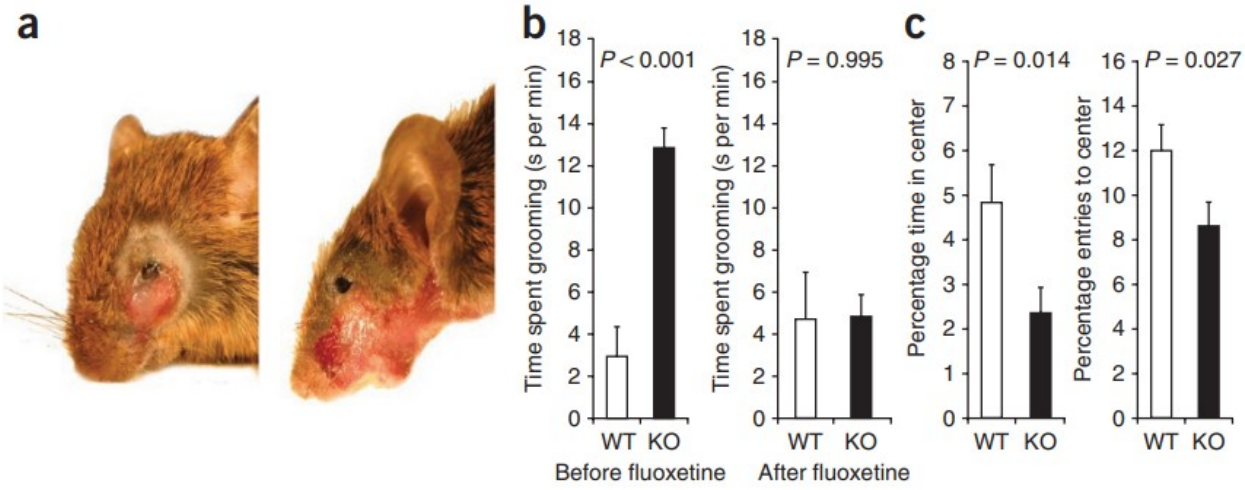
Ting and Feng presented the Sapap3 null-mice experiment as a promising step toward understanding the neural foundations of OCD-like behaviors, represented by the observed excessive grooming in the subjected animals and the elevated anxiety levels. The authors suggest that the absence of the Sapap3 gene disrupts the CSTC circuitry, which according to them plays a fundamental role in regulating OCD emergence and manifestation. Besides the OCD-like behaviors which were interpreted as a direct evidence for OCD emergence, the partial efficacy of serotonin reuptake inhibitors in alleviating some of the symptoms was also suggested to provide further relevance of the model. This being said, it is important to note the limitations acknowledged in the study. For instance, not all genetically modified mice responded to the reintroduced Sapap3 gene, and significant behaviors persisted in many cases. Additionally, while the absence of Sapap3 in the striatum was associated with defects in synaptic communication, similar issues were not observed in the hippocampus. Considering this, the authors presented several more studies, in order to strengthen their claim that genetic animal models are providing a great value in understanding complex human psychiatric conditions.

Slitrk5 null mice – S. Shmelkov et al., 2010

The second study that G. Feng and J. Ting considered promising for uncovering the precise causal factors of OCD is the work of Sergey Shmelkov et al. (2010) – *"Slitrk5 deficiency impairs corticostratial circuitry and leads to obsessive-compulisve-like behaviors in mice"*. Very similar to the Sapap3 knockout mice, the researchers generated Slitrk5 null mice, hypothesizing that abnormal expression of this gene may lead to specific behavioral patterns associated with OCD-like behaviors (S. Shmelkov et. al, 2010). By replacing Slitrk5 with lacZ gene which served as a marker, Shmelkov and colleagues found that Slitrk5 is highly expressed throughout the entire central nervous system.

After several generations of Slitrk5 knockout mice were produced, researchers began their behavioral analysis. Excessive grooming behavior was observed in all genetically modified animals, which have led to extremely severe wounds and skin lesions. Additionally, open-field and elevated-plus-maze tests were performed, in order to measure their anxiety levels. Similar to Sapap3 null mice, Slitrk5 mice also have shown less time spent in the center compartment and a lower number

of entries into the center compartment in the open-field test, as well as spending reduced amount of time in open arms within the elevated-plus-maze test (Shmelkov et. al, 2010). Interestingly, the authors also found that Slitrk5 null mice had significantly reduced size of the striatum, indicating that the Slitrk5 gene could have a developmental role in specific brain regions.



The image shows the consequences of the excessive grooming in Slitrk5 knockout mice, as well the data obtained from the two tests performed to measure anxiety-like behaviors. For a detailed information, please see Sergey Shmelkov et al. (2010) – *"Slitrk5 deficiency impairs corticostratial circuitry and leads to obsessive-compulisve-like behaviors in mice"*.

Hoxb8 null mice

In 2002, another study made in the Capecchi lab, found yet another cause for compulsive grooming in mice. Mutant Hoxb8 mice not only have shown excessive grooming and fur loss, but this was also followed with 100% penetrance (Capecchi, 2002). This behavior even extended beyond self-grooming to include grooming also of wildtype cage-mates. What made the results of the study so surprising is the fact that normally Hoxb8 is a member of a large family transcription factors, responsible for body patterning during embryonic development (Ting and Feng, 2011).

An important development regarding Hoxb8 was reached in 2010, where it was revealed that Hoxb8 originates from bone marrow-derived microglia that migrates into the brain postnatally (Schlegelmilch et al., 2010). The researchers have found that bone marrow transplantation from mutant mice to their wildtype relatives has led to increased grooming and fur loss, whereas transplantation of bone marrow from wildtype mice into Hobx8 null-mice had immense positive

effects, stopping fully the compulsive self-grooming in a mutant mice subset. The authors argue that such findings underscore an important departure from the conventional understanding of obsessive-compulsive disorder as a purely neuronal disorder situated within the cortico-striatal-thalamic-cortical (CSTC) circuits. The consideration that Hoxb8 originates from non-neuronal cells and only postnatally migrates to the brain via bone marrow poses reasons to reevaluate the CSTC neuropathological paradigm.

EAAC1/Slc1a1 null mice

The last study discussed by Feng and Ting is related to the role a protein called EAAC1 plays within the brain, and the Slc1a1 gene which encodes EAAC1. In a nutshell, this protein has a double function – it helps control the spread of a chemical called glutamate, and it is also involved in the transportation of another substance, cysteine, responsible for making an important antioxidant called glutathoine (Aoyama, et al., 2006).

When EAAC1 gene deletion was performed, mutant mice exhibited various brain problems because of the increased sensitivity to oxidative stress, a condition which damages brain cells via harmful molecules due to the decrease of glutathoine production. Additionally, obscure behavioral anomalies occurred after one year of age, such as increased aggression, excessive self-grooming and a loss of fur. Since EAAC1 is also expressed in the CSTC circuitry, the authors suggest that such evidence supports the OCD-CSTC paradigm. To further emphasize on the potential of EAAC1 to be involved in obsessive-compulsive disorders, Ting and Feng also mentioned that a very rare variant of Slc1a1 has been identified in a single OCD family to date (Wang, et al., 2009).

Conclusions and Discussion by Jonathan T. Ting and Guoping Feng (2011)

The message with which the authors conclude their summary of findings is straightforward and unambiguous – we should celebrate genetic animal models, invest more resources and time in developing such, and expand our experimentation over non-human species, in order to facilitate greater understanding of the human nature. What stands alone, however, besides this clear message, is the somewhat contradictory admission of the authors that, citing (page 842–848):

“No animal model will adequately capture all aspects of a complex human disorder…”

Before diving into a detailed reflection on the implications of these findings and their actual reliability and relevance to the human psychiatric condition of OCD, it is essential to first outline and consider the premises, assumptions, motivations, and conclusions presented by Ting and Feng. This step is important because scientific language can sometimes give the impression that claims and assumptions are established facts, even though in reality they are just marking the starting point from where the scientific investigation begins. This have many times led to experiments losing their whole relevance if the initial assumption proves to be wrong, so with this in mind, let us now take a closer look at the arguments and points that Ting and Feng raised.

- The concept of Animal Modeling – this is the central premise around which all findings, suggestions, claims, assumptions, results, and experiments in the carefully described above paper gravitate. In its very essence, the idea is that animals can serve as an adequate model of human cognition, behavior, and biology, which makes them a suitable replacement of humans for experimental purposes, especially regarding experiments that could normally have major health-related concerns and/or ethical consequences. The main reason for this replacement, therefore, is the belief that results obtained from experiments with animals can unfold something about the human nature. In a narrower perspective, the authors hold the belief that humans and mice genetics are similar enough so that mice can adequately and sufficiently represent the genome of human beings, at least when it comes to OCD. Moreover, this also implies that OCD can be causally linked to the same specific genes both in mice and humans.

- Invasive experimentation – Ting and Feng unambiguously state already in the third sentence of their paper that, citing: *"The ability to perform invasive techniques in genetic animal models of OCD will be crucial for rapid advances in this field…"*. Since gene modification can be achieved only by invasive methods, such a statement implicitly posits that OCD animal modeling could be best achieved through such an approach.

- Motivation related to uncovering *"the precise causal factors"* of OCD – the authors state several times across their paper that what precisely causes OCD is not known. For this they assume multiple reasons. For instance, that there was little attention paid to the promising cortico-striatal-thalamo-cortical (CSTC) circuitry in the brain that indicates linkage to repetitive behaviors, or, citing: "*… the ethical issues that preclude the majority of invasive*

procedures in human patients and the lack of well-established alternative animal models for exploring such detailed mechanistic insights".

- CSTC as theoretical framework – the authors select the cortico-stratial-thalamo-cortical circuitry as their theoretical foundation, given that previous non-invasive imaging attempts with OCD patients have identified abnormal patterns of activity in this circuit. In addition, other findings related to the basal ganglia circuitry that have shown partially closed loops through the CSTC pathway, have lead to the hypothesis that OCD symptoms emerge out of hyperactivity in the orbitofrontal-subcortical loops caused by a disruption in the activity of the opposing ganglia pathways (Saxena and Rauch, 2000). It is important to mention that Ting and Feng explicitly define this hypothesis as an "*oversimplification of the neural circuitry that gates obsessions and compulsive behaviors*", nevertheless, they believe that it still remains useful for devising testable hypotheses.

- Assumption that OCD causes can be reduced to single genes and/or brain circuits – perhaps the biggest assumption in the above described paper is that obsessive-compulsive disorder occurs and/or remains in the patients due to specific genetic anomalies. Additionally, these genes can be traced back to the CSTC pathway and their main function is mainly related to this specific circuitry.

- Assumption that excessive self-grooming is an analog to human compulsions – as most of the papers discussed in the overview of Ting and Feng have shown, the main behavior identified as an analog for human compulsion is self-grooming, sometimes to the point of penetration and open wounds.

2.4 First Case Study Discussion – Critical Analysis and Reflection

This part of my thesis will be devoted to a thorough reflection, critical analysis, and discussion of the methodological conclusions drawn by Ting and Feng. As the overarching focus of my work revolves around the challenges posed by animal modeling in cognitive science research, the central point of this exploration will revolve around the two key tools science provides to test the reliability and validity of such methodologies, namely, external and internal validation. Through an examination of how Ting and Feng's study aligns with these validation principles, I aim to critically reflect on and analyze their approach, with the goal to provide a perspective for

understanding the broader methodological implications of their work. This reflection will address some of the prime challenges of using animal models, especially in translating findings to complex human mental conditions.

External Validation

I shall begin with the bigger picture before breaking down the specifics of the discussed publication. To put it bluntly, one might ask the question – How can we translate findings from animal experimentation to humans? After all, aren't humans too different from mice, to reliably assume that whatever we observe in these animals can be extrapolated to us? Also, could obsessive-compulsive disorders actually occur in mice, given the cognitive, emotional, and behavioral complexity associated with this mental condition? Such questions, indeed, are extremely relevant and I am not the first person who poses them. In fact, more and more researchers acknowledge the shortcomings of animal experimentation. For example, as I have mentioned in the External Validity section, Pandora Pound and Merel Ritskes-Hoitinga (2018) have stated that species differences will always undermine external validation. Such a view is supported by the fact that the failure of drug trials to project beneficial effects for humans is to a large extend attributed to the inability of animal models to predict clinical efficacy (Geerts, 2009; Perrin, 2014; Wall and Shani, 2008) and safety (Waring et al., 2015; Hwang et al., 2016). Interestingly, although such evidence appears to be acknowledged by most scientists, predominantly their attention is directed towards improving internal validity measures, whereas the problem of external validation seems to be neglected (Pound and Ritskes-Hoitinga 2018).

To adequately address the external validation problem with respect to the selected case study of Ting and Feng, I shall first describe what exactly is being extrapolated and how the findings translate to humans. To begin with, one can say with confidence that all four papers selected by the authors as promising try to recreate obsessive-compulsive disorder in mice and then translate their findings to humans. However, in order to be more precise, I should say that in the four papers only OCD-like symptoms were induced, and if I need to be exact – not a single comment or claim regarding obsessions was made throughout the paper/s, and only the behavior-like aspects of OCD were presumably addressed. From an experimental perspective this makes a lot of sense, as there is literally no way to track, tell, predict, or know the mental content of a mice, or for that matter the mental content of any given animal. After all, in the taxonomy of OCD, obsessions are classified as "intrusive thoughts". From an external validation perspective however, this is not only a problem in

the paper of Ting and Feng, but a general issue of any study that uses animal models to tell something about obsessive-compulsive disorders. It is too obvious to ignore that the method of modeling animals is simply insufficient when it comes to investigating obsessions. In fact, Ting and Feng seemingly acknowledge this by saying the following in the discussion part of their paper (page 842–848):

> "...no single model is sufficient to recapitulate the entirety of the human condition of OCD. This is especially true given that the presence and content of obsessional thinking that is intimately tied to compulsive behavior is perhaps impossible to assess in a mouse."

One does not need a thorough interpretation to understand this message. On the other hand, knowing that such a fundamental problem exists makes the extrapolation of results from mice to human OCD patients extremely questionable for several reasons. First, in most formal descriptions of obsessive-compulsive disorders, the repetitive behaviors (compulsions) are interpreted as a kind of coping mechanism that a person with OCD exhibits to combat their own intrusive thoughts (Robbins et al., 2019). In fact, compulsions are so intimately linked with the mental content of the patients, that the type of thoughts can ultimately determine their behavior. For example, the different subtypes of OCD (i.e. contamination or hoarding OCD) are basically based on the type of intrusive thoughts one has. Considering this, an individual who has obsessions that almost all objects around are dirty, polluted, and contain harmful and dangerous microorganisms, will act very differently to an individual who is always doubtful about the completeness of their actions in terms of locking the door, or turning off the oven. In the first case, the behavior often includes excessive hand washing, where in the second excessive checking of the object and returning to it over and over again. Although one cannot claim with ease that compulsions are only secondary to obsessions, it is more than obvious that intrusive thoughts are fundamental for understanding OCD. If no animal model is capable of capturing the obsessive aspect of OCD, then how can we be sure that any emerging behavior which is categorized as compulsive has anything to do with OCD?

Another issue that arises from an external validation point of view is the degree of OCD generalization presupposed in those animal modeling studies. Not only they assume that this mental condition can be adequately represented just by mere repetitive behavior, but the type of behavior also does not need to change regardless of the interventions made. To put it bluntly, the fact that animals have undergone gene modification/deletion of completely different proteins such as Sapap3, Hoxb8, or Slitrk5 and exhibited the same OCD-like symptom described as excessive

grooming, seems to be of no or little importance for Ting and Feng. Although humans exhibit quite different behavioral patterns depending on the type of OCD/intrusive thoughts one has, it seems that for a good/promising model of OCD in mice it is not expected that different types of repetitive actions should emerge. If I need to reflect a little more on that, an obvious question arise – can an animal model even represent a ***type*** of OCD? If there is no possibility to investigate and track in any way the obsessional thinking in a mouse, then is it possible to develop a model that addresses specifically contamination OCD, or checking OCD? Based on my research, and certainly not within the discussed case study, there is no animal model that even tries to model a specific type of obsessive-compulsive disorder.

In some of the selected studies by Ting and Feng (i.e. M. Capecchi et al., 2002) excessive grooming was reported to be very similar to the human condition of Thrichotillomania. The parallel was made due to the similarity in the behavioral output, as Thrichotillomania is characterized by the recurrent, irresistible urge to pull out one's own hair, leading to noticeable hair loss. Individuals with this condition often engage in such a behavior as a way to cope with stress, anxiety, or other negative emotions (D. Woods & D. Houghton, 2014). Interestingly, it seems that even in this case researchers try to bind the excessive self-grooming in mice to some kind of human compulsions only after they observe the repetitive actions that emerged from the animal, showing that there was no intentionality for a particular outcome, nor for targeting Trichotollomania specifically when the model was developed. But even if we are about to ignore the so far mentioned external validation issues and accept the results and the assumptions around them, it becomes extremely problematic when one tries to assess what causes, say, Thrichotillomania. Based on the collection of promising papers the authors proposed, it is not even clear whether such a condition comes from the brain or not. For example, Capecchi et al. (2002), found that Hoxb8 originates from bone marrow, a fact that does not even fit within the underlying neurophatological CSTC framework, as the Hoxb8 gene is only introduced to the brain and does not emerge from it.

The over-generalization issue has another important aspect that should be considered, namely, the cognitive dimension. Throughout the paper of Ting and Feng, no differentiation between the cognitive functionalities involved in different types of OCD was made. The limitations of animal models in accurately representing the variations of cognitive dysfunctions in different Obsessive-Compulsive Disorder subtypes become especially pronounced when considering the specific brain areas implicated in these cognitive functions. For instance, contamination OCD often involves cognition such as heightened attention and perception along with extremely strong focus

towards potentially polluted surfaces. Those functions often are associated with the parietal and occipital lobes (Corbetta & Shulman, 2002). Conversely, checking OCD is closely linked to working memory and decision-making processes, primarily regulated by the prefrontal cortex. Furthermore, it is important to mention that the intricate neural circuitry of OCD extends beyond these examples, involving regions such as the amygdala which is responsible for emotional regulation. Considering this, it becomes clear why models such as Sapap3 are greatly impaired and unable to capture the complexity OCD has to offer even on a neural level alone. While the striatum is undoubtedly implicated in some aspects of OCD, particularly in the context of compulsive rituals and behaviors due to its association with the reward system and habit formation (M. Malavaez & K. Wassum, 2018), it represents just a fraction of the broader neural networks that are involved in this disorder. A similar problem occurs when one tries to limit the field of exploration to the CSTC circuitry. Although this network plays an important role in a broad range of cognitive functionalities, those functions are by no means limited to it. An exclusive focus on one region or neural circuit undermines the ability to comprehensively replicate diverse cognitive dysfunctions seen in different OCD subtypes, which ultimately reduces the model's translational relevance to the human condition.

When we pay close attention to the main method used in all presented studies, namely, invasive models and genetic modifications, some confounding elements emerge which could also have an impact on the generalization and extrapolation of the results from mice to humans. First, in all models presented in this case study, the genetic deletion has wiped the selected gene from the entire organism, not only from the associated brain areas and circuits. There is a fundamental assumption made by all of the researchers, which is that if a brain region has low density of a given gene or that gene is not "*highly expressed*", then the relevance of this gene for this particular region or circuit diminishes. This is best shown in the Sapap3 paper (2007, p. 4), where the authors state:

> "We further tested whether the synaptic defects are unique to the striatum by examining synaptic transmission in the hippocampus, in which Sapap1, 3 and 4 are highly expressed. We found no defects in CA3–CA1 hippocampal basal synaptic transmission in Sapap32/2 mice, ***probably reflecting functional redundancy or compensation by other members*** of the SAPAP family. Together, these findings identify an important role for SAPAP3 in postsynaptic glutamatergic synaptic function at cortico-striatal synapses."

With this statement the authors completely neglected the possibility that Sapap3 plays an important role in the hippocampus. Moreover, they suppose that the function Sapap3 plays in the hippocampus could be simply “redundant” given the other Sapap members also expressed there. It is quite intriguing that this is the only sentence in the whole Sapap3 paper that pays attention to the hippocampus, especially given that most of the behavioral tests designed to measure anxiety were related to navigating through open fields and mazes. As it is well described in the neuroscientific literature, the hippocampus is a vital structure nestled within the brain’s medial temporal lobe and it is highly involved in various cognitive functions, especially spatial memory and navigation (Moser et al., 2008). With the discovery of the so-called place cells, the role of the hippocampus in encoding and retrieval of spatial information is considered crucial. Given that the Sapap3 protein is found to be highly expressed in that area, there is a possibility that the orientation functions were damaged due to its deletion. Such prospects could be the reason why Sapap3 null mice spent less time in the middle of the open field. As we know, in the paper this behavior was labeled as if the animals have shown “anxiety” where in reality they might just have been unable to properly navigate themselves.

What is interesting, however, is that the authors also mentioned the compensation phenomenon. This is especially important from an external validation point of view, as genetic deletions have been shown to trigger unexpected and uncontrolled genetic compensations. Scientific inquiries, exemplified in studies such as M. El-Brolosy et al. (2017), reveal instances where the removal of one gene prompts compensatory adaptations in others, resulting in an unpredictable sequence of genetic interactions. A phenomenon with very similar compensatory characteristics that functions in an analogous way is the so-called neuroplasticity. For example, blind individuals demonstrate an astonishing ability to adapt to their environment by altering the visual cortex to accommodate non-visual functions such as braille reading (Sadato et al, 1996). Such evidence poses serious question on whether any compensatory mechanisms were triggered during the presented animal models and whether the observed behavioral changes such as excessive grooming and heightened anxiety were due to genetic and developmental compensations or because of the intervention made.

Another example can be given in the case of Slitrk5 null mice, where researchers found a significantly reduced striatum volume. This have made the researchers suggest a relevance of the protein to this brain region and its potential developmental role. However, when they replaced the Slitrk5 gene with another marker gene (LacZ), they have found that Slitrk5 was actually expressed

throughout the entire central nervous system. This not only goes beyond the CSTC circuitry, but also shows the importance of understanding the role of a given gene in the organism as a whole. To elaborate, it is not so obvious why Slitrk5 has developmental attributes only when it comes to the striatum, since no other brain regions were found to have a reduced volume, although the gene is expressed throughout the whole brain. Failing to answer such questions is failing to understand what is the exact role and functions of the deleted gene. Considering this, the claims that this gene plays role particularly in compulsive behaviors such as excessive grooming, seem to be rather bold and unproven, as the exact function of Slitrk5 is well understood neither in mice nor in humans.

By critically examining the paper of Ting and Feng on the neurobiology of obsessive-compulsive disorder (OCD), from an external validation point of view some substantial concerns are evident, undermining the translational relevance of the proposed OCD animal models. First, no matter exact procedure regarding the proposed models, the difference between mice and humans, especially on environmental, sensory, and emotional level, are simply too significant to reliably attempt generalizations to humans who suffer from OCD (Pound and Ritskes-Hoitinga 2018). Second, the inability of any model to capture obsessional thinking and the cognitive dimensions involved in this mental condition further hinders the translation from mice to humans. Moreover, the over-generalization of behaviors across different genetic interventions, the exclusive focus on specific brain regions and/or circuits, as well as the lack of consideration for compensatory mechanisms additionally compound the external validation issues. However, the problems just described are not yet touching the internal validation components. Therefore, I will now pay a closer look to those, in an attempt to critically evaluate the relevance of the experimental design and how supportive such can be in uncovering the very nature of OCD.

Internal Validation

In order to look into the robustness of the studies presented by Ting and Feng from an internal validity point of view, we must take a closer look into the experimental design chosen to measure the effect of single genes over mice behavior. However, since several studies used the same measurement tests, especially on a behavioral level (i.e. open field, dark-light emergence, and elevated zero maze tests), the description of those will be thoroughly presented just once, in order to avoid redundancy and repetition.

To begin with, in the experiments performed with Sapap3 null mice, the animals were subjected to a series of a highly invasive procedures, on top of the genetic deletion that was done to their parents prior birth. A mere week after their birth, all mice were anesthetized and immobilized in a mouse head holder, a process that introduces significant stress to the animals and potential complications. While being restrained, the mice were then subjected to a procedure involving direct injections of GFP lentivirus into the striatum. This process was conducted through both sides at eight locations per hemisphere. On top of that, all mice received acute sagittal brain slices, approximately 300 micrometers thick, further emphasizing the physical toll and immense room for complications those animals have gone through just to be capable of participating in the experimental design. The brain slices were performed so that a monopolar stimulation electrode can be placed in the corpus callosum, and a field recording electrode in the dorso-lateral striatum. Such procedures can be extremely hazardous, as they have a high chance of introducing unwanted tissue damage, which can lead to developmental problems and symptoms. In fact, looking purely at the methodology where brain slices and striatum injections are performed, one can easily understand why similar research approaches are strictly prohibited with human subjects, and even patients. Yet, the question from an internal validation perspective is – to what extent can such procedures introduce the risk of confounding factors to occur? And the answer to such a question is very simple – invasive procedures have an extremely high chance of causing unexpected and uncontrolled complications, thus increasing substantially the risk of confounding factors. To put it bluntly, this is the main reason we perform brain slices for experimental purposes on animals and not on humans.

Let us now consider an example regarding what types of complications could occur when intracranial surgeries are performed. In their paper *"Cognitive and behavioural effects of chronic stimulation of the subthalamic nucleus in patients with Parkinson's disease"*, Daniele et al. (2003, page 179) describe the effects of deep brain stimulation (DBS) on Parkinson patients. The study presents twenty individuals who were subjected to a bilateral stereotactic subthalamic nucleus implant, in order to try to relieve them from the tremors which were unaffected by alternative treatments (i.e. drugs). Although the post-surgical analysis of the motor control of the patients have shown a great improvement over the ability of the subjects to control their tremors, the authors also have presented the so-called "adverse events". They write:

> "Eleven of the 20 patients experienced adverse events. Four had persistent complications: psychic akinesia (one patient), mild worsening of dysarthria (two patients), and a mild worsening of hypophonia (one patient). The patient who showed worsening of

hypophonia also experienced transient ballistic or choreic dyskinesias of the neck or limbs, as did one of the patients who showed a worsening of dysarthria. Seven further patients had only transient adverse effects: three showed transient manic symptoms with hypersexuality, one experienced paraesthesiae, and two experienced paraesthesiae and diplopia. One of these latter patients had a perioperative complication, namely a mild subarachnoid haemorrhage, the only presenting symptom being transient diplopia. No permanent lesions occurred in any patient, as confirmed by neurological examination and magnetic resonance imaging.“

As we can see, more than half of the patients had negative, unexpected, and unwanted effects after the brain surgery. Additionally, three of the patients had manic symptoms with hypersexuality. This is particularly important to note, as manic symptomatic can lead to impulsive behaviors that are also involved in obsessive-compulsive disorders. In other words, complications and adverse events can be not only physical (i.e. bleeding), but also cognitive. In fact, The National Institute of Health in United States (2023) have published a comprehensive analysis regarding the adverse events emerging from neurosurgery procedures with human patients. In their report, the numbers display a dark picture where every fifth human (one in five; 1/5) who was a subject of brain surgery, had at least one adverse event 30 days after the procedure. Although a roughly 20% chance for an AE to occur is scary, the chances of dogs for experiencing adverse effects from intracranial surgeries is even higher. A study conducted in 2018 by Forward et al., shows that early postoperative neurological deterioration was seen in 45% of the subjects, and half of them also had postoperative complications. Considering, on the one hand, that no specific measurements for preventing unwanted adverse events related to the neurosurgical preparation of the animals were noted, and on the other, that the complication rate in such procedures is extremely high, there is almost a guarantee, at least from a statistical perspective, for the occurrence of adverse events within the subjects that took part in the Sapap3 experiment.

From an internal validation perspective, the inclusion and endorsement of one’s own experiment as a crucial evidence to support another scientific claim, introduces a clear issue of bias. Guoping Feng was one of the main authors behind the Sapap3 null-mice experiment. He is also the main author of our case-study paper **“***Neurobiology of obsessive–compulsive disorder: insights into neural circuitry dysfunction through mouse genetics***”,** in which the claim that invasive animal models are crucial for understanding the nature of human obsessive-compulsive disorders was made. The act of self-citing and evaluating one’s own work introduces a serious risk of

confirmation bias, expressed in the possibility that the researcher may selectively emphasize or interpret results to align with their preconceived hypotheses. Decisions such as this one can be found as compromising the empirical reliability and objectivity of the scientific method. A great example of research bias is given by Ioannidis (2005), where he highlights that biases such as selective reporting and interpretation are a significant contributor to the prevalence of false-positive findings in scientific literature. Having this in mind, the choice of G. Feng to rely on his own experiment as a pivotal part in his argument illustrates a major problem from an internal validation standpoint.

An important fact that connects all the presented studies by Ting and Feng, despite their variety in approaches in terms of targeted genes, invasive procedures, and selection of tests, is the presence of observational coding related to the self-grooming behavior exhibited by the animals. To put it differently, in all studies the key variable supposed to signal for the development of obsessive-compulsive disorder are the excessive self-cleaning acts. It is important to mention that humans suffering from OCD often report an awareness of their compulsions and experience distress, frustration, and/or negative emotions from the lack of control over their rituals (Macy A. et al, 2013). Considering this, it is rather important that such aspects are taken into account when a model is created, in order to eliminate potential confounding factors. Unfortunately, none of the studies reported specific metrics that are implemented to measure emotional states, except the tests which are supposed to address anxiety. However, since the anxiety tests were all performed separately from the behavioral coding of compulsive behavior, it is not clear whether negative emotional states and distress were experienced from the animals when they groomed themselves. This opens the door for further debates, for example, whether the animals felt an actual pleasure from the self-grooming. Such questions could become of great importance, especially when we consider that, for example, in the Hoxb8 knockout study 100% of the mice were reported to exhibit excessive grooming. Given that all animals developed severe or mild lesions which can be tremendously painful, it is possible that a sense of pleasure pushed the animals beyond the pain in a continuous self-harming. For example, Poon & Thompson (2018) have described reward-related neural mechanisms that correlate with thoughts of nonsuicidal self-injury. Considering this, it is not even necessary that a feeling of pleasure is present, but it could also be that due to the genetic manipulation, the neural mechanisms responsible for processing reward-related stimuli were also affected.

In summary, the discussed internal validation arguments display some considerable shortcomings occurring in the studies introduced by Ting and Feng. First, the potential of numerous confounding factors to occur was highlighted with respect to the selected procedures for experimental conduct. For example, the choice for invasive procedures such as brain slices and electrode insertions are bringing high chances for complications, as well as unforeseen and unwanted adverse events, which ultimately can influence the observable behavior and cognition of the animals. Additionally, a potentially biased assessment was pointed out, signaling a lack of internal validation measurements when it comes to evaluating a study for promising or not. Then, a discussion related to perhaps the most emphasized on behavior, the compulsive self-grooming, considered the possibility of misinterpretation of the animal's actions which could further weaken the translational value of the presented scientific work. Overall, there are several internal validation gaps, especially if we consider the design principles that should be followed for ensuring a robust scientific rigor when one decides to deploy animal models.

One could be easily tempted to try and provide a general explanation regarding the reasons behind the emergence of so many challenges and limitations, when one tries to model a complex human mental condition in an animal. For example, an argument could be that although mice are closer biologically to humans than, say reptiles, they are still considerably different than us especially when it comes to the development and emergence of mental disorders. Another argument might be that conditions such as OCD are fundamentally grounded in cultural and social dynamics, therefore the richness and differences in these domains between mice and humans are the causes for so many extrapolation limitations to emerge. Therefore, with my next choice for a case study I will try to partially address, and also challenge, some of these explanations. To further reduce the differences between humans and the modeled animal, I have selected a primate animal model, as there is no other species that is genetically and biologically closer to us. Moreover, as the emergence of complex mental disorders could also be problematic due to the involvement of several cognitive processes intersecting with culture and social dynamics, the next case study should mainly focus on one cognitive ability, which in our case will be – language. With this in mind, I will now continue to the second case study - **"***The neural and cognitive correlates of aimed throwing in chimpanzees: a magnetic resonance image and behavioural study on a unique form of social tool use***".**

3 Second Case Study – Modeling Primates for Studying Language Evolution

Paper Name: "*The neural and cognitive correlates of aimed throwing in chimpanzees: a magnetic resonance image and behavioural study on a unique form of social tool use***"**

Authors: William D. Hopkins, Jamie L. Russell, and Jennifer A. Schaeffer

Year of Publishing: 2012

Journal, impact factor: Philosophical Transactions of the Royal Society B: Biological Sciences (Volume 367, Issue 1585, Impact factor 5.0)

Language is perhaps the single most advanced cognitive aspect that sets us apart from other species. The ability not only to communicate to each other, but also to transmit knowledge, experiences, ideas, and concepts across generations, is arguably the most important reason for our superiority as a species. Yet, it is still a mystery how exactly language emerged in such a sophisticated way, that could allow eventually to evolve in alternative forms, such as writing and even braille. From a scientific perspective, one of the intuitive and preferred ways to approach such a question, naturally involves studying means of communication. After all, the use of language already presumes the presence of communication. However, from my perspective there are some notable pitfalls with this approach. For instance, the concept of communication is by far the more broader compared to language, involving not only structured verbal and/or written language, but also visual, auditory, and tactile cues that aim to transmit and communicate information. Considering this, one should be aware that if a study aims to investigate language evolution through communication, it can easily go beyond the borders of the concept of language itself, especially when non-verbal forms are factored. If that is the case, then one must also acknowledge the prime and underlying assumption – that language emergence is intimately linked with non-verbal forms of communication.

William D. Hopkins, Jamie L. Russell, and Jennifer A. Schaeffer decided to approach the question of language evolution through creating a chimpanzee model based on a social tool use – a form of communication evident in some primates. More specifically, they focused on aimed throwing – an ability that suggests the presence of some complex cognitive functions, such as planning (Suddendorf & Corballis 2007). As it will become evident later on, the authors of the

paper *"The neural and cognitive correlates of aimed throwing in chimpanzees: a magnetic resonance image and behavioural study on a unique form of social tool use"* (2012, page 30) suggest that the evolutionary selective increase of the ability to aim and throw *"may have refined the neural architecture of the left hemisphere so as to eventually support other complex motor sequencing actions, including language and speech"*. As this research proposes an entirely different approach to animal modeling than the transgenic mice studies, in the next paragraphs I will delve into the details of this study, presenting how animal models are used to uncover evolutionary insights. After that, a thorough analysis and a discussion based on the findings, assumptions, and conclusions made by the authors will follow.

3.1 Experimental Design and Deployment of the Model

To pursue a better understanding of the evolution of language, the authors created a complex experimental design that consist of several stages. First, magnetic resonance imaging (MRI) was employed to examine the inferior frontal gyrus (IFG) which is considered as homologous of the Broca's area in humans – a specific brain region associated with language and speech production. Additionally, the motor-hand area also known as KNOB was in focus. After the scans were obtained, image processing and data collection related to the brain regions of interest were performed. Following the MRI scans and image processing, a behavioral experimentation was then deployed, where a modified version of the primate cognition test battery (PCTB) was selected as a measurement tool to assess various aspects of cognitive abilities (Herrmann et al., 2007). The cognitive dimensions covered by this battery include social cognition tasks, causality (i.e. cause-and-effect relationships), physical cognition tasks (spatial memory, ability to differentiate between quantities, etc.), and theory of mind tasks.

Starting with magnetic resonance images, a sample of 78 chimpanzees were scanned, of which 24 were males and 54 females. The animals were anywhere from 6 years of age to 51 years, where the average age was 23.05 (s.d. = 11.80). For the in vivo scanning, subjects were first immobilized by ketamine injection (10 mg kg^{-1}) and subsequently anaesthetized with propofol (40–60 mg (kg^{h-1}) $^{-1}$) following standard procedures (Hopkins et al., 2012). After that, the animals were transported to the MRI facility and remained under anaesthesia for the whole duration, usually between 40 to 60 minutes. As the study goal was to investigate the relationship between aimed throwing and language evolution, 38 of the subjected animals were able to throw reliably, and 38 couldn't. It is important to mention that the chimps who could reliably throw were taught to do it

through reward-related conditioning. However, the type of reward is unfortunately not explicitly mentioned.

When it comes to the image processing and data collection, the researchers measured the inferior frontal gyrus (IFG) separately for the left and right hemispheres of the brain. Following a procedure described in (Hopkins et al., 2008), they defined the total area and size of IFG by looking at specific features in the scan, such as the precentral inferior (PCI) and fronto-orbital sulci (FO). The authors performed this procedure for each hemisphere, in order to compare the total volume of IFG consequently in the right and left sides of the brain. Similarly, the Motor-hand area (KNOB) was quantified separately for both hemispheres, where the dorsal and ventral edges of the KNOB along the central sulcus (CS) served as markers (Hopkins et al., 2012). Finally, the researchers analyzed the composition of different brain tissues and their ratios. More specific, they compared the amount of white matter (WM) to grey matter (GM), in order to assess potential differences between both hemispheres.

In terms of behavioral testing, the authors decided to deploy the Primate Cognition Test Battery (PCTB) originally developed by Herrman et al. in 2007. The battery aims to measure different cognitive dimensions, including physical and social cognition. However, it is important to mention that some of the original aspects in this battery were removed and not performed, while others were modified. The researchers justify this by arguing that such an approach will allow them to, citing *"better address the questions at hand, given the past experience and environmental constrains of our subjects"*. Additionally, they also mention that most of the subjects were tested alone, however, there were animals that felt uncomfortable being separated from their group. Therefore, for these specific individuals the tests were performed with one other conspecific, with whom they felt comfortable. Unfortunately, the number of chimps who felt uncomfortable is not mentioned by the authors.

Physical Cognition Testing

With respect to physical cognition, several tasks were performed including spatial memory task (three trials), object permanence (nine trials), causality visual (six trials), tool properties (six trials), rotation (nine trials), transposition (nine trials), relative numbers (thirteen trials), and causality noise (six trials). For the spatial memory testing, the animals were tasked with remembering the locations of hidden food rewards. The food was placed under two from total of

three opaque cups that were turned upside down on a test table in front of the subjects. If a chimp successfully located both rewards, the trial was considered as successful. The task related to testing object permanence expanded on that, investigating the ability of the animals to follow a food reward after it was displaced to a new location that was explicitly manipulated. Trials were considered successful if the chimp successfully found the hidden food item without searching in locations that were not manipulated.

When it comes to the rotation task, a food reward was placed on a platform, after which the platform was turned to 180 or 360 degrees. To be considered as a successful trial, the animal should locate the food reward in the correct location from the first time after the rotation was performed. In a similar manner to the previous tasks, the transposition experiment was designed in a way that a food reward is hidden in one from total of three possible locations and then the baited location is changed in three different ways. First, the baited location is simply switched with one of the unbaited ones. Second, after the switch of the baited with unbaited location, the two unbaited locations were also switched. Lastly, the baited location was switched with one of the unbaited ones, and then with the second as well. Success was achieved if the animals correctly traced the reward and choose the baited location.

The relative numbers trials examined the numerical discrimination of the subjects, presenting them with pairs of plates containing different quantities of food pieces. Individuals were tasked with choosing the plate with the larger quantity of food. If correctly doing so, the trial was considered successful.

The testing of noise and visual causality comprehension was performed in two different setups. For noise causality, the animals had to point out one of two containers that had a food reward inside. Before pointing out, both containers were lifted and shaken, where the full one (with the reward inside) made a noise, where the empty one did not. Consequently, for the visual causality a food reward was placed underneath one of two boards lying flat on a test table. As the bait made one of the flat boards to tilt a bit, a success was considered when the chimp correctly points out to the tilted board.

The final physical cognition task examined tool use. Subjects had the choice between tools with varying effectiveness in obtaining food rewards. They encountered different scenarios where they had to select which tool could successfully retrieve the reward, based on its physical

properties. Notably, the authors state that three tool property tasks were not included from the original PCTB collection, namely the “Bridge”, “Broken Wool”, and “Tray Circle”.

Social Cognition Testing

When it comes to social cognition, researchers used several approaches from the PCTB collection, namely, comprehension (six trials), production (four trials), attentional state (eight trials), and gaze following (three trials) tasks. The comprehension task was modified compared to the original Herrmann et al. work, where the authors argue that this modification would bring more value to their experimentation. In a nutshell, the task was initially designed to measure abilities to comprehend communicative signals by implementing the same table-and-cup set-up described in some of the physical cognition tasks above. However, Hopkins et al., changed the task by not including the “Mark” condition (Herrmann et al., 2007), and shifting the originally described position of the subject on the left and right sides of the enclosure. Then, by using a gaze or a combination of a gaze with manual point, the subjected animals were tested to showcase their understanding of those communicative signals. On the contrary, the production task evaluated chimpanzees’ ability to produce communicative signals. The subjects observed how one of the experimenters baited a location with a food reward, after which a second experimenter entered the enclosure. A successful trial was considered when the animal correctly indicated the baited location by pointing it out to the second experimenter.

In the attentional state task, the researchers tried to alter the attentional cues of the apes in various ways, such as standing with a face and body turned towards or away from the subject. Success required the animal to use different communicative signals that interact with the attentional state of the experimenter. This ranged from tactile and visual cues when the experimenter was facing the chimp, to auditory signals (i.e. banging the cage) when the experimenter’s attention was diverted. It is important to mention that Hopkins et al., (2012) added another condition, where they conducted the different attentional states in a more familiar to the animal setting, namely, the usual testing table that was used for previous tasks before. The authors note that the same requirements regarding modality-specific communication were kept.

Results & Interpretation

To analyze all acquired behavioral data, Hopkins et al. (2012) averaged the proportion of correct trials over each of the twelve tasks. The computation was then performed for five of the basic cognitive dimensions described by Hermann et al. including Communication (Comprehension, Production, and Attentional State tasks), Causality (Visual and Noise Causality tasks, as well as Tool Properties), Space (Spatial Memory, Object Permanence, Rotation, and Transposition tasks), Quantity (Relative Numbers task), and Theory of Mind (Gaze alternation task).

The study initially compared the proportion of white matter (WM) to gray matter (GM) in the IFG and KNOB areas between male and female chimpanzees, reporting that no significant differences were found with respect to sex differences. After that, they explored the differences between animals who could reliably throw (THROW+) versus animals who couldn't (THROW-). Significant distinctions were found in THROW+ chimps, showcasing a higher WM to GM proportion in the KNOB area. Additionally, researchers also found significant differences between left- and right-handed throwers, where the right-handed animals had greater leftward asymmetries in both IFG and KNOB brain areas.

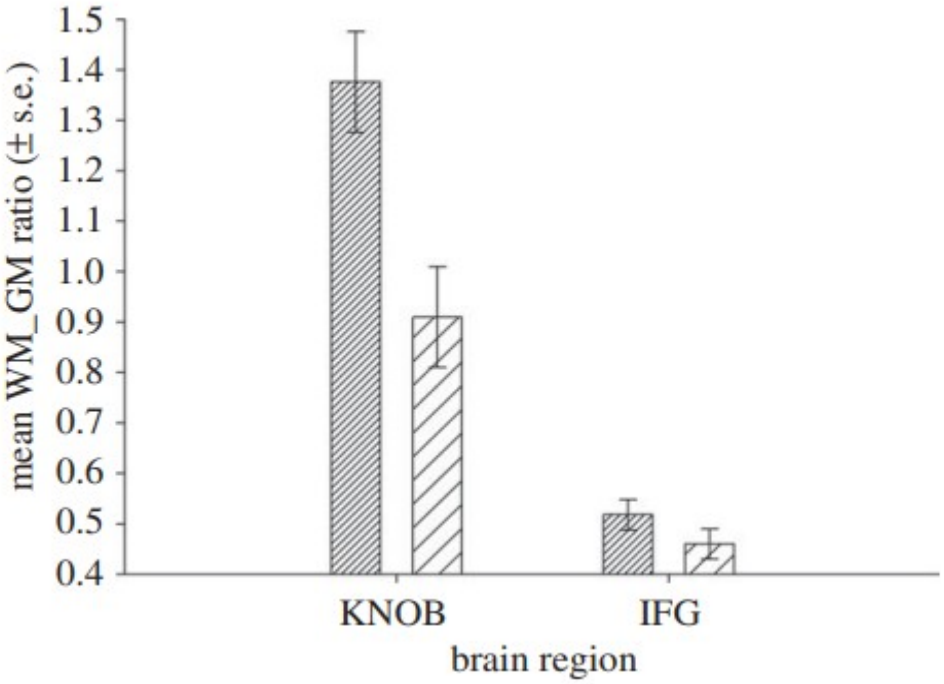


Fig. X, Hopkins et al., (2012). The graph displays the mean white-to-grey matter ratio in both KNOB and IFG brain areas. The narrow-striped bars indicate THROW+ chimpanzees, where the wide-striped bars indicate THROW- chimpanzees.

When it comes to the behavioral analysis, the study analyzed 91 chimpanzees in their performance on the Primate Cognition Test Battery (PCTB) of which 39 consistently threw objects, and 52 did not. A multiple analysis of variance (MANOVA) was used for comparison of the mean PCTB between THROW+ and THROW- groups, showing a significant main effect for the throwing group. After that, F-Tests were performed to analyze the specific differences between the two animal groups for each cognitive domain. Researchers concluded that chimpanzees who can reliably throw outperformed those who cannot in the communication domain.

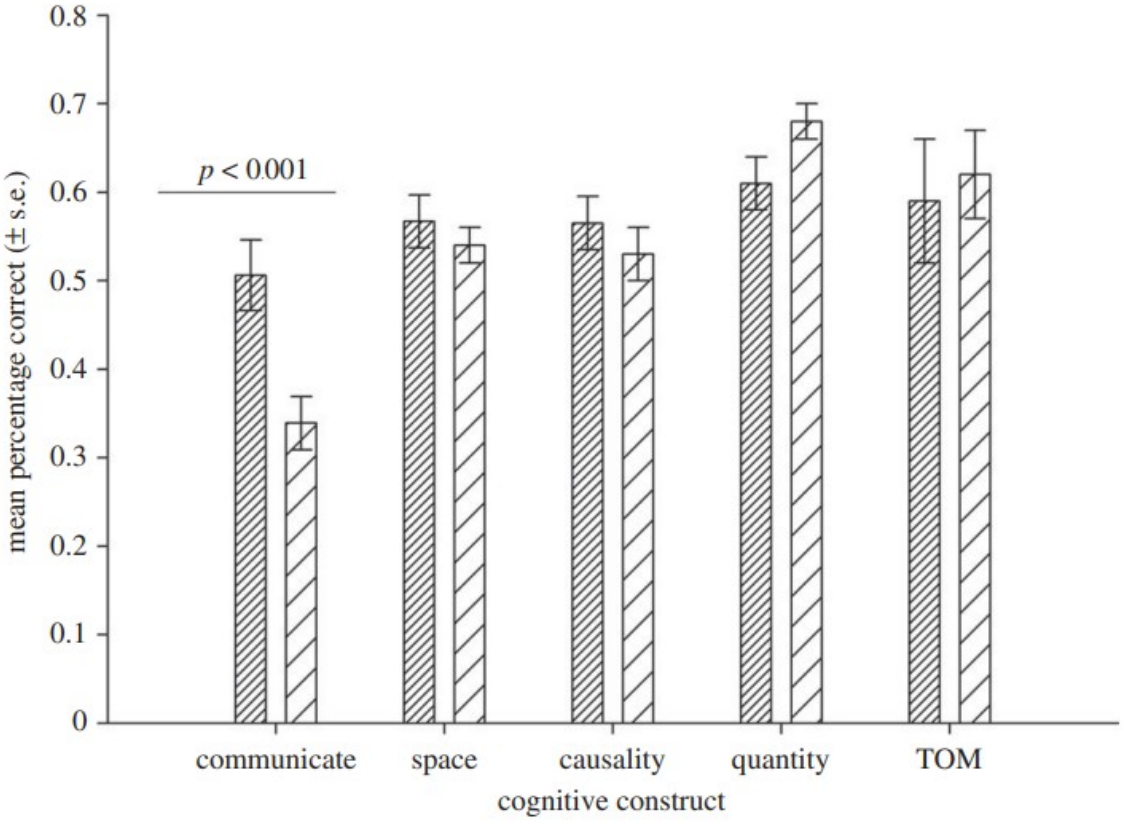


Fig. X, Hopkins et al. (2012). The figure represents the mean percentage for the five cognitive dimensions tested with the PCTB tasks. Narrow-striped bars indicate THROW+ chimpanzees, where wide-striped bars indicate THROW-.

Based on the results discussed above, the authors conclude that throwing skills in chimpanzees are correlated with increased cortical connectivity in the Broca's area, as well as in the motor-hand (KNOB) brain region of the precentral gyrus. Given that the presence of higher WM to GM ratio is an indication for more myelinated interneurons, the authors suggest that learning to throw could impact the connectivity between premotor and primary motor cortex in the brain. When it comes to the asymmetries reported between left- and right-handed throwers, the authors conclude that such results are consistent with the evolutionary hypothesis that "throwing may have served as a preadaptation for the neural adaptation of motor programmes necessary for complex motor actions, including language and speech" (Calvin, 1983).

Based on the results obtained from the PCTB testing, the authors also suggested that chimps who learned to reliably throw are better at communication tasks than their counterparts. Furthermore, Hopkins et al. (2012, page 45) argue that such results should be considered as an explicit association between the cognitive foundations for throwing and the ability to engage in successful intraspecies communication. To expand on this conclusion, the researchers state that, citing:

> "...the development and acquisition of throwing skills by chimpanzees operates in a manner similar to the emergence of manual gestural communication. As noted previously, the motivation for throwing in chimpanzees is largely to alter the behaviour of other individuals (be it human visitors or conspecifics). For this reason, the apes that have learned to throw have acquired an ability to understand how their behaviour affects the behaviours of others. If the same individuals apply these basic skills within the context of (i) understanding gestures, (ii) producing gestures, and (iii) using attention-getting behaviours when a human experimenter is inattentive to them, then it would follow that the THROW+ individuals should outperform the THROW- apes on the PCTB test. Moreover, these same skills may not offer any advantage in tasks that are not communicative in function."

To acknowledge limitations, Hopkins et al. discuss that the positive association between white-to-gray matter ratio and throwing should be further investigated, as it could be that chimps

who could reliably throw inherently had higher white-to-gray matter ratio, allowing them to learn this skill, instead of the assumption that the learning process itself resulted in increase of white-to-gray matter. However, despite these limitations, the authors believe that their findings are a solid evidence for a potential evolutionary link between complex motor actions like throwing and the development of language and speech.

In a nutshell, the study suggests that the ability to reliably aim and throw in chimpanzees correlates with enhanced neural connectivity in regions associated with language and speech, indicating that throwing may have served as a preadaptation for the neural structures involved in complex motor actions, including language development. In a way, the authors suggest that such a connection highlights the potential evolutionary link between specific motor skills and the emergence of language.

3.2 Second Case Study Discussion – Critical Analysis and Reflection

The goal of uncovering the mysteries that surround language and speech evolution in humans is a formidable task. Science offers many methods through which one could approach such an endeavor – from deploying animal models and comparative biological studies, to computational models that simulate diverse evolutionary pathways and weights a variety of factors, suggesting potential causes of language emergence. In their paper, Hopkins et al. (2012) decided to tackle this challenge through modeling chimpanzees, as they are some of the closest biologically species to us, humans. In the next paragraphs, I will share my view regarding their experimental design and critically discuss whether the selected … from the authors approach, and the conclusions they made, are supported from internal and external validation point of view. Before doing so, however, it is important to bring to our awareness the main assumptions and hypotheses that are underlying the whole research:

- One of the main assumptions of the study is that there is a strong connection between non-verbal communication and language/speech emergence, suggesting that non-verbal communication is a prerequisite for language/speech and it lays the foundation for such.

- Assumption that the inferior frontal gyrus (IFG) brain area in chimpanzees is an analogue to humans Broca’s area, and therefore involved in language/speech production and communication.

- There is a suggestion that higher white matter in specific brain areas results in higher socio-cognitive performance. For example, the authors propose that higher white-to-gray matter ratio in the IFG and KNOB areas of chimpanzees is intimately linked with their communication abilities.

- The authors modified the main test battery for behavioral and cognitive testing by removing specific tests and manipulating the initial setup of other selected experiments. This assumes that the introduced changes will still result in obtaining relevant data, as no information regarding a validation of the performed changes in the test was reported.

- It is also assumed that the primate cognition test battery and its communication dimension are offering an appropriate research environment to study language and speech emergence.

External Validity

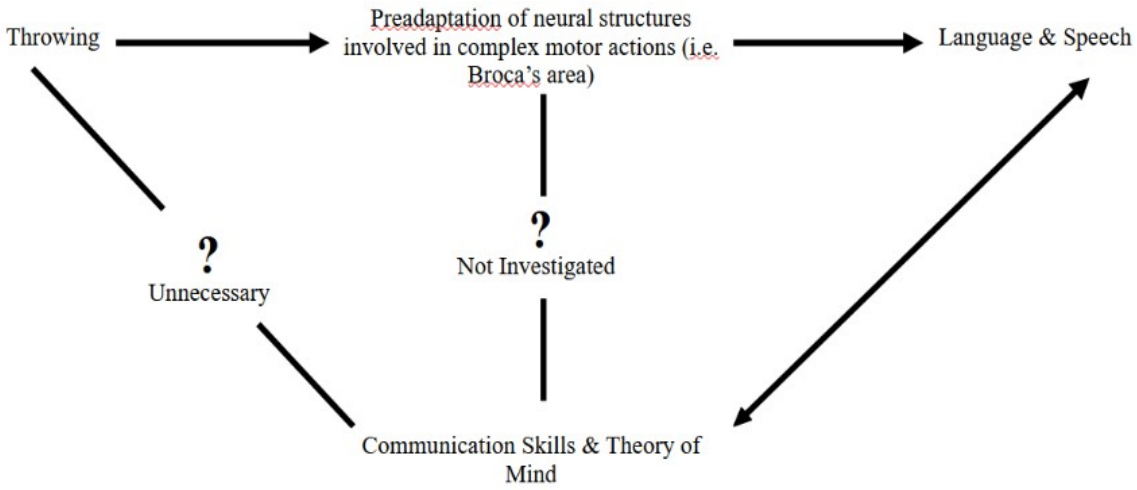
Taking an external validity perspective, one must first and foremost ask oneself what exactly in this research is being extrapolated from chimpanzees to humans, and after identifying that - how transferable are the results of this study. The first question can be answered relatively easy – the study claims that aimed throwing contributed to the development of language and speech in humans. More specifically, it shaped and adapted the brain regions that are currently considered as crucial for language and speech production. On the other hand, it is not so straightforward answering the transferability question, although from an animal modeling perspective, it is the more important one. In fact, there are several issues that arise when one tires to extrapolate the findings from these chimpanzees to the evolution of human language.

To begin with, it is important to distinguish between the non-verbal communicative signals exhibited by the animals through social tool use, and the language and speech production that is unique to humans. That is, even if we assume that aimed throwing could have contributed to the development of more sophisticated communication strategies and behaviors, this does not necessarily mean that language and speech are among these strategies and behaviors. That is because non-verbal communication might include facial expressions, gesturing, and signaling through bodily movements, whereas language and speech are directly connected with vocalizations, semantic processing, and syntax. If we take a closer look at the social cognition tasks deployed in

the study, it becomes apparent that in every communication-related trial, a success is considered only when some sort of non-verbal communicative signaling is exhibited. However, I believe that the lack of any measurement of vocalization abilities of the animals considerably limits the relationship of the findings to speech and language. While non-verbal signals, such as gesturing, can indicate communication attempts, their extrapolation to human language, and in particular to human speech, is significantly more restricted compared to vocal signals. In fact, vocalizations are fundamental for speech production and are an essential part of any language. For instance, key language structures, such as syntax, are expressed verbally through vocal sounds. Therefore, I believe that investigating vocal behaviors and communication in chimpanzees can provide much closer parallel to human speech and language than focusing solely on the non-verbal gestures expressed by the animals. It is also important to mention that from a neurological perspective, the main focus of the study was related to the equivalent of the Broca's area in chimps, namely, the KNOB and the inferior frontal gyrus (IFG). In humans, Broca's area is perhaps the most heavily involved brain region in speech production and semantic processing (Davis et al., 2008). This further emphasizes the need of taking into account the functional aspects of these brain regions when one deploys a language-based animal model, instead of merely relying on structural similarities.

Another problem that occurs from an external validation point of view is the fact that there is no way in which the evolutionary claim made by the authors can be empirically confirmed, nor rejected. Evolution-based events are spanning over vast amounts of time, often stretching for thousands, if not millions of years. The factors that could have shaped or tilted an evolutionary process in one direction or another are therefore boundless, making it an extremely challenging task to pinpoint the exact elements that ultimately drove the change. Considering this, even if the study had shown that there is no activation of communication specific regions during the act of aimed throwing, this would not necessarily negate the possibility that such an action might have influenced speech and language evolution in humans. On the contrary, showing a strong activation in the IFG/KNOB areas of a chimp during aimed throwing would not necessarily translate into the evolution of speech production in humans. With this in mind, I believe that the second part of the study, namely, the behavioral testing, is completely redundant and does not contribute to the main evolutionary hypothesis/claim made by the authors. To be more specific, the authors are making two distinct arguments. From one side, they argue that "throwing may have served as a preadaptation for the neural adaptation of motor programmes necessary for complex motor actions, including language and speech" (Calvin, 1983; Hopkins et al., 2012). From another side, the

researchers are trying to demonstrate a correlation between proficient throwing abilities and enhanced communication skills in chimpanzees.



In essence, the behavioral experimentation does not address the main argument – that aimed throwing preadapted the IFG and KNOB areas, so that language and speech can emerge. In fact, the only evidence that the PCTB trials are actually providing is that higher connectivity in the IFG and KNOB areas result in both - better communication AND better throwing skills. The researchers themselves admit that whether the ability to throw has resulted in higher connectivity, or the presence of higher connectivity allowed the animals to learn how to throw, is still a question. In a way, one cannot draw a causal link between throwing and communication abilities in chimpanzees. For example, it might be that more intelligent chimps are simply better at both throwing and communication tasks, rather than throwing itself positively impacting communication skills.

One of the most important aspects of external validity is the concept of ecological relevance. In a nutshell, this notion represents the degree to which the conditions examined in a given experiment accurately represent the natural environment and behaviors of the species being studied. When examining evolutionary processes, accounting for ecological relevance is particularly important. The reason for that is simple – evolutionary changes and adaptations emerge mainly in response to selective pressures occurring in the natural habitat of the organism. Therefore, in order to draw conclusions about evolutionary phenomena, having an experimental setup that is as close as possible to the natural environment of the subject should be a logical requirement.

Now, let us consider the experiment conducted by Hopkins et al. (2012). Starting from the subjects, the authors explicitly state that all chimpanzees that were used for this study were

“members of a captive colony housed at YNPRC in Atlanta, GA, USA”. Additionally, the primate cognition test battery was performed in a captive environment, where the setup of the tests was everything but representing natural conditions (i.e. having a table and a cup for the Comprehension task). Such an approach introduces serious translational challenges, especially when evolutionary conclusions are drawn based on entirely artificial surroundings and captive populations. To begin with, the individual development, range of behaviors, and natural abilities of captive chimpanzees might be greatly altered and/or constrained due to the unnatural settings and housing in which they live. The variety of stimuli, the sociological context, and the evolutionary pressures in a natural environment can offer much more diversity and complexity that could directly influence the function of a given behavior such as aimed throwing. Moreover, chimpanzees that live in a captive environment could have different motivations to engage in aimed throwing, and/or could exhibit such a behavior in different contexts compared to wild populations.

To illustrate some of the issues I have raised, let us consider the example for social tool use provided by the authors. They discuss how captive apes point to otherwise unattainable foods in the presence of humans, thus manipulating the human to get the food for them. However, such type of social dynamics and environmental factors could not be relevant in the natural settings of the animals. It might be that pointing, or throwing, serve completely different functions for wild chimpanzees compared to their captive relatives. For instance, Wilke et al. (2022) reported that wild chimpanzees may use referential showing gestures (such as pointing) to direct others’ attention to objects simply for the sake of sharing.

To put it simply, if the behaviors observed in captive animals are not representative of those exhibited by their wild relatives, then evolutionary processes cannot be reliably reflected by captive research. In other words, since the aimed throwing exhibited by captive chimpanzees was influenced by artificial factors such as habitat limitations and/or social disruptions, then linking it to the evolution of human language becomes problematic. Considering that natural selection is perhaps the prime mechanism of evolution, relying on experiments where “natural” is missing presents a puzzling contradiction, to say the least.

Internal validity

I shall start this internal validity section with an emphasis that the research conducted by Hopkins et al. (2012, page 39) had two main phases, namely, the MRI scanning which was

performed to measure the white-to-gray matter ratio in the IFG and KNOB brain areas, and the behavioral testing deployed to evaluate specific cognitive abilities of the subjected animals. I would also like to bring the reader's attention to the numbers reported by the authors. When it comes to the magnetic resonance imaging, they state:

> "Magnetic resonance images were obtained from **sample of 78 chimpanzees**, including **24 males and 54 females**. The subjects ranged in age from 6 to 51 years (mean = 23.05, s.d. = 11.80). All the chimpanzees were members of a captive colony housed at YNPRC in Atlanta, GA, USA. Within the sample, there were 38 chimpanzees that reliably threw, and these individuals were matched on the basis of sex, age and scanning protocol with **38 chimpanzees that did not reliably throw**."

We can see that a total of 78 chimpanzees had undergone MRI scanning, but only 76 of them were matched based on sex, age, and scanning protocol (38 animals who can throw + 38 who cannot = 76 animals in total). There is absolutely no information in the whole paper regarding what happened to the two chimps that were not matched. From a scientific perspective, this could be a major concern as many questions that have the potential to compromise the study arise. For example, what happened with the results of the two animals? What is the reasoning of them being excluded – was it that they created outliers (i.e. extreme difference in age/throwing abilities), or that they failed the scanning protocol? Since the number of female subjects (54 animals) is twice as much as the number of male subjects (24 animals), how were they actually matched in two groups of 38 animals each? Additionally, considering purely the age variation of the group (from 6 to 51 years of age, with a standard deviation of 11.80 years), there is no information whether age was taken into account as a covariate with respect to cognitive abilities. This could also be of great importance, as it might be that due to limited developmental time and accumulated experience, older subjects are simply better throwers than younger ones, or they have more matured brain development that could lead to higher white-to-gray matter ratio. Unfortunately, all of these questions were not addressed whatsoever, where the only suggestion that some actions were taken to account for internal validation, is the sentence (page 39):

> "...Whitin the sample, there 38 chimpanzees that reliably threw, and these individuals were matched on the basis of sex, age, and scanning protocol with 38 chimpanzees that did not reliably throw".

Considering that the numbers of subjected animals reported in the first phase of the study are already worrying (i.e. two subjects potentially excluded for unknown reasons), and that there is no information of how the two groups were matched on key criteria such as age (a key factor for cognitive and brain development), one should be careful when assessing the second phase of the study, namely, the behavioral testing. Moreover, since there is a direct dependence between the two phases, such as the argument that the ability to throw results in a higher connectivity in language specific brain areas and thus leading to better communication skills, it should be crucial that the behavioral testing was performed on the same population that undergone the MRI scanning. This being said, let us take now a closer look on what the authors actually reported (page 43):

> "As noted earlier, our laboratory has administered the PCTB test to **91 chimpanzees**, and within this sample, there were **39 apes that threw consistently (THROW+) and 52 that did not (THROW-)**. In this next analysis, we compared THROW+ and THROW- individuals in their performance on the PCTB task."

The discrepancy between the number of subjected animals in phase one (78 chimpanzees, of which 2 chimpanzees were excluded for unknown reasons, 38 could reliably throw and 38 couldn't) and phase two (91 tested chimpanzees, of which 39 could reliably throw and 52 couldn't) is puzzling, to say the least. To begin with, there is a considerable difference between the sample sizes of MRI scanned animals versus behaviorally tested animals. Given that the paper does not report on why there is such a discrepancy and how it emerged, there are several possible scenarios. First, it could be that some of the chimpanzees that went through MRI scans were also used for behavioral testing. However, if this is the case then it is crucial to understand the selection criteria for the additional chimpanzees. Why there are additional subjects in the behavioral testing and how they compare to the animals that were part of phase one? Were they younger or older, did they also undergone MRI testing in other research, did they differ in cognitive and/or physical abilities in any way? All of these questions have a direct impact on the results of the study and could potentially tilt the outcomes in one direction or another. Given that a total of 15 animals were added to the behavioral part of the study (13, if we assume that the two 2 chimps that remained unassigned to a group in the MRI scans also participated), the absence of any information regarding the selection criteria seriously hinders comparability between phase one and phase two, while also raises concerns about sample selection bias.

Of course, we should not forget that the scenario outlined above is just a possible one. In fact, the study does not provide any information suggesting that the population which had undergone MRI scanning was also included in the behavioral testing. There is also another possibility, namely, that the two samples of chimpanzees are simply independent. This would simply imply that a direct comparison between the two groups cannot be made. For instance, if the group that undergone behavioral testing was not scanned, it would be impossible to assess whether the chimpanzees that could reliably throw and/or performed better at cognitive tasks had a higher white-to-gray matter ratio. In other words, without overlapping subjects, the researchers cannot conclusively link MRI findings to behavioral outcomes. Additionally, since there is no information about the second group in terms of age and sex of the subjects, there is yet another confounding variable that remains unaccounted.

Continuing with the subject of internal validity, we should not neglect another fundamental aspect to this experiment, namely, the concept of throwing. Throughout the whole paper, the authors speak about aimed throwing, chimpanzees who could reliably throw, and/or such that could not reliably throw. However, there is a clear lack of explicit explanation regarding how throwing ability was actually assessed. Given that the differentiation between throwing groups is critical for both phases of this study, the absence of any definition how subjects were assigned to one group or another poses some serious problems from a scientific perspective. First, without knowing what exactly constitutes “reliable” throwing, it is quite challenging to understand what was the skill gap between the chimpanzees and how they were classified as throwers or non-throwers. Second, replicating this experiment would be a challenging endeavor, considering the absence of precise assessment criteria of throwers versus non-throwers, which is necessary to ensure consistency in further research. For instance, were the primates categorized based on their ability to hit a target while throwing, did the distance of the throwing matter, did it matter if the throwing behavior was exhibited in a social context? The lack of transparency regarding how reliably to assess the throwing ability of the animals also signals for potential risks related to subjective biases during the assessment process. Finally, it should be mentioned that given the unclear and potentially flawed assignment of chimpanzees to throwing or non-throwing groups, all observed differences in cognitive performance and/or brain structure might have been mistakenly linked to throwing ability, instead of other unaccounted variables. Such obscurity not only undermines the translational value of this research, but also brings into question whether the obtained results are actually addressing the main hypothesis of the study.

In summary, the conduct of the experiment and the conclusions made by Hopkins et al. (2012) raise several concerns from both external and internal validity perspectives. When it comes to external validity, extrapolating non-vocal communicative signals to human language and speech is quite the oversimplification, to say the least. The lack of measurements of vocalization abilities in chimpanzees clearly undermines the transferability of the findings, especially if we consider that an argument about the evolution of human speech was made. Moreover, given that evolutionary processes span over vast amounts of time, the behavioral testing can neither confirm nor refute whether a preadaptation of the IFG and KNOB areas emerged as a result of aimed throwing. In a sense, the second phase of the study simply does not support the evolutionary claims made by the authors. Finally, the concept of ecological relevance was completely neglected during the experimental conduct. All research activities were conducted with captive animals in artificial settings. This is problematic for several reasons. First, captive conditions could significantly alter the natural behavior and development of the animals. Second, all evolutionary processes related to chimpanzees, or to any other species for that matter, unfolded as a result of the selective pressures of their natural habitat. Omitting to account for ecologically relevant factors further weakens the external validation of the study.

The study's internal validity is also compromised by notable methodological limitations, particularly the discrepancy in the number of animals across the two phases: 78 in the MRI phase and 91 in the behavioral phase. This raises concerns about a lack of overlap or independence between the groups, making direct comparisons problematic or impossible. Additionally, the criteria for categorizing chimpanzees into throwing and non-throwing groups are unclear, introducing potential bias in group assignments and limiting future replication. Furthermore, details on matching animals based on sex, age, and scanning protocols are insufficient, especially given the uneven distribution of male and female animals and throwers versus non-throwers. Finally, the exclusion of two chimpanzees from the MRI scans raises questions about their role in the whole study, as well as for their omission from the results.

To conclude, I believe that the overall external and internal validity of the study is unreliable, and at best – seriously limited. The methodological mistakes I have outlined above raise concerns whether the results actually represent the dependent variables (i.e. white-to-gray matter ratio). This alone makes the study potentially inconsistent already within its own borders, not to mention if these results are to be interpreted for other chimpanzee populations or even extrapolated to humans. However, even if the internal conduct was robust and factors such as sample size and

control for bias were addressed, this would still not be enough to reliably account for all external validation aspects, which would allow us to resiliently translate the findings to humans.

4. Discussion

In this final chapter of my thesis, I shall present the main conclusions and reflection points related to the two case studies I have presented, as well as outline the arguments which underscore the methodological issues emerging when animal models are deployed for advancing our understanding of human cognition. I am completely aware that a generalization over all cognitive animal models based on these two case studies cannot be made. However, I strongly believe that most of the issues I found within the two discussed case studies are fundamental, meaning, they are not bound to the selected studies only, but rather are valid considerations for animal modeling experimental research in general. The primary reason for such a conviction stems from the fact that although the two selected studies are both deploying animals to model human cognition, the approach to this deployment is substantially different. First, humans and chimpanzees are both coming from the mammalian family of primates, as opposed to mice, which are coming from the family of rodents. This should, at least in theory, enhance the extrapolation value that we could obtain from chimpanzees, as from an evolutionary perspective we share much closer ancestors and biological similarities. By focusing on chimpanzees and not exclusively on mice, I have come as close as possible to us, humans, if we are to consider that species differences is a major factor that can influence the plausibility of the final translation. Second, the experimental approach in both studies follows completely different trajectory – in the OCD mice model, invasive methods such as brain surgeries, electrode implantation, and diverse types of genetic modifications were performed. On the contrary, the chimpanzee language model did not include any invasive intervention, but rather focused primarily on the behavioral aspects demonstrated by the animals. However, despite these essential differences, I believe that there is a significant overlap when it comes to the limitations that arise from both case studies.

The first common limitation that becomes evident in both studies is emerging from the juxtaposition of genetic similarity and cognitive complexity. There is already a substantial amount of evidence suggesting that understanding complex cognitive phenomena, like language and speech or obsessive-compulsive disorders, demands far more than genetic similarity – it requires a multi-dimensional and interdisciplinary approach that would allow us to encompass the different layers of entanglement. From my perspective, there is no doubt that the genetic proximity between chimps

and humans was the prime reason behind the decision to use them as a language and speech model, instead of, say lions, who also have an immensely rich social behavior and utilize diverse vocalizations. The same argument can be also applied to the selection of mice for the development of OCD models – we share similar genes, therefore mice seems to be a good candidate for mimicking the genetic makeup of psychiatric disorders. Yet, the crucial question such an approach fundamentally fails to address is – Can language and speech, or obsessive-compulsive disorders, be reduced to genetic and biological proximity? And the answer to this question is – No, they cannot, as otherwise this would imply that the environment of the organism, the individual differences and experiences, the social and cultural context, and even the considerably richer neurological complexity that humans have, are simply irrelevant factors to the involved cognitive phenomena unfold. Having said that, both studies have demonstrated strong tendency towards biological reductionism, where all of the above mentioned external validity factors were simply not considered.

The second common issue that emerge from both case studies is the obvious limitation to primarily rely on behavioral analogies, in order to draw conclusions about human cognition. If we consider the chimpanzee language model, one could confidently say that the evolution of human language and speech was approximated to a series of tasks, that essentially captured communication instead of language, and had nothing to do with speech and syntax – both of which fundamental to human language. Similarly to that, the mice OCD models presented by Ting and Feng were mainly focused on behaviors that seemingly mimic obsessions, such as excessive self-grooming. Yet, compulsions were completely omitted, although some researchers believe that there is a causal link established between compulsions and obsessions, where obsessions are emerging to serve as stress and anxiety coping mechanisms (Robbins et al., 2019). Both examples I just provided are underpinned by the same problem – reliance on behaviorally based analogies is simply destined to oversimplify the investigated phenomena, excluding interrelated mental processes that ultimately shape the way these phenomena are exhibited in humans. On top of that, it is extremely difficult, if not impossible, to scientifically prove that the analogy one relies on, is a reliable and accurate representation of the real phenomenon of interest. If we ask, for example, "Is the exhibited excessive self-grooming truly an obsession?" – we cannot say with certainty either yes, it is, or no, it is not. It might resemble an obsession, but we have no way to confirm if it actually is. Consider now, that most of the conclusions drawn from this research are based precisely on the assumption, that this behavior is indeed an obsession. In view of this, I believe that the ultimate goal of science

should focus on testing assumptions to establish solid facts, rather than drawing conclusions based on assumptions.

Another important aspect that I would like to highlight in this section, is the considerable challenge of establishing robust and reliable internal validation measures within the context of animal modeling. After carefully reflecting on both case studies, it becomes more and more clear to me that as the scope of achieving external relevance expands, the task of maintaining an adequate internal validity framework grows in complexity. In other words, the further one wants to extrapolate the results across different populations, environments, species, and multi-layered cognitive phenomena, the more challenging is to deploy experimental controls that can ensure meaningful results. To put this into context, let us consider the case study of transgenic OCD mice models. The ambition to replicate human-like obsessive behaviors in mice and simultaneously link them to specific genes, introduces numerous obstacles which an internal validation framework should control for. Whether the issue is that genetic modifications can only be applied to the whole organism rather than to specific brain regions or neural circuits, or that for a precise measurement of specific neuronal activity one must perform a highly hazardous brain surgery and implant electrodes which by itself introduces immense stress to the subjected animal, it becomes increasingly difficult to discern if the observed behaviors genuinely represent human OCD or are simply a by-product of the intervention (i.e. increased stress response and/or damage in the brain after the surgery, or unaccounted neural circuits affected by the genetic modification). In a similar way, the ambition behind the chimpanzee language model is to uncover the evolutionary roots of a cognitive ability present only in humans – language and speech. If we are to search for a rigor in this particular case, it is not even clear how one could reliably model such a phenomenon without introducing any bias or inconsistency. The need to introduce modifications in the Primate Cognition Test Battery which was selected as a measurement tool, as well as the lack of an explicit and detailed criteria for assessing the throwing abilities of the chimpanzees, are simply showcasing some of the experimental design problems that emerge as a result of a significantly broaden external scope. Considering the arguments above, I believe that the further an animal model is stretched to approximate human cognition, the more susceptible it becomes to internal validation issues. In a way, an interesting paradox emerges – the more an animal model tries to achieve a greater human relevance, the more it undermines the very experimental rigor necessary, to serve as a reliable scientific method.

Being conscious that my following argument would be much more of a speculation, than an established fact, I still find it relevant and beneficial to be expressed within the realm of this thesis. My intuition is that one of the main issues with cognitive animal models, namely the more and more recognized failure to achieve a credible external validity, is to a large extend similar to the replication crisis in the field of psychology. In the past several decades, many psychological experiments have drawn conclusions about the human psyche based on the so-called WEIRD experimental populations – Western, Educated, Industrialized, Rich, and Democratic (Henrich et al., 2010). This introduced a substantial amount of psychological research that falls short of uncovering anything fundamental about our psychological processes. Instead, the resulting replication crisis simply highlights how a whole generation of scientists can overlook the importance of external validation by not considering enough the importance of human diversity. For instance, large-scale collaboration between psychology researchers across the globe have shown that only 39% of psychological science papers published in Nature have been successfully reproduced (An open, large-scale, collaborative effort to estimate the reproducibility of psychological science, 2012). Given that research conducted with humans, which also aims to uncover more about humans, have a significant chance to miss the mark if external validation is not considered, it comes to me rather intuitive that cognitive animal models which do not account for the same validity criteria are simply bound to fail. In fact, Pound and Ritskes-Hoitinga (2018) report that the non-reproducibility problem could be even higher when it comes to modeling animals, ranging between 51% to 89% failed attempts in biomedical research. At some point, all of the above mentioned points are simply begging the question – If there is so much evidence suggesting that modeling animals is a method that is significantly limited in terms of reliability and reproducibility, then what exactly justifies subjecting animals to experiments that involve substantial pain, suffering, or death? And even more – If these animal models fall short of achieving their intended goals on a large scale, should we not question the very ethics of continuing such practices?

4.1 Ethical implications of using animal models

Throughout my thesis so far, I have deliberately refrained from discussing the ethical dimensions related to experimenting with animals. The reason for that is not because the experiments I referred to do not involve substantial suffering, nor because I don’t find such ethical considerations relevant. Quite the opposite, I believe that scientists, regardless of their field of expertise, should be actively engaged in continuous discussions about the ethical aspects related to animals used for experimental purposes. However, the ethical debate involves many perspectives

and layers, and it simply deserves a thesis of its own. Having said that, my choice to focus precisely on the methodological limitations of using animal models was intentional. This is because I believe that if the ultimate goal of any scientific endeavor is to acquire robust and reliable knowledge that can serve as a foundation for further advancement of our understanding, then using animals to model humans has proven to be not only extremely inefficient, but also a generally ineffective and unreliable method. Considering this, it is impossible to deny the ethical relevance of such a statement, as perhaps the main argument that is used to justify animal suffering in experimental setups is that the perceived value of the research (i.e. finding the root causes of OCD) outweighs the moral cost of inflicted harm and pain (DeGrazia, 1999). However, I simply believe that such an argument essentially collapses when animal experimentation fails to produce consistent and reliable insights.

You may notice that when I speak about ethical problems, I am not specifying whether the experimentation is using animals for modeling purposes, nor do I differentiate whether the deployment is within the field of cognitive science or biomedical research. This is because, from an ethical perspective, using animals for achieving experimental objectives is already problematic even without these specifications. There is an extensive literature with respect to the ethical debate about using animals in experiments (Paixao & Schramm, 1999; Ferrari, 2020) that highlights different welfare-related aspects. While I have so far tried to maintain a neutral voice in describing the procedures animals endure, a quick glance through the past sections of my thesis will vividly expose the practical nature of why people bother to argue over welfare issues. Just as an example, let us consider the OCD mice model – raising the animals with genetic modifications that subject them to uncontrollable and serious self-harm, slicing their brains to prepare them for the experimental objectives, performing direct injections in the head, and killing them in a way that the brain literally freezes, are just some of the procedural examples that shine forth. And if one dares to ask – Why is all of this suffering caused? The answer usually is – Because the expected benefit is holding a higher value than the harm caused (DeGrazia, 1999, Grimm et al., 2019). Within the animals welfare debate, this question, and the answer, are having somewhat central position. More specific, the main direction of this ethical conversation revolves around advocating for practices that reduce, refine, or replace animal use wherever possible (Russell & Burch, 1959), improving housing conditions and minimizing suffering, as well as a call for a political shift in how animal experimentation is justified (Ferrari, 2020). However, while the welfare-oriented arguments are important, I believe that there is another crucial dimension to the ethical debate that seems to be rather overlooked. To be precise, scholars such as Regan (1986) have questioned the very ethics that

permit the instrumental use of animals in the first place. As the author writes in his book “The Case for Animal Rights”, page 1:

> “What's wrong - fundamentally wrong - with the way animals are treated isn't the details that vary from case to case. It's the whole system. The fundamental wrong is the system that allows us to view animals as our resources, here for us-to be eaten, or surgically manipulated, or exploited for sport or money. Once we accept this view of animals - as our resources - the rest is as predictable as it is regrettable. Why worry about their loneliness, their pain, their death? Since animals exist for us, to benefit us in one way or another, what harms them really doesn't matter - or matters only if it starts to bother us…”

Welfare standards, while being of a crucial importance, do not examine the very essence of using sentient beings for experimental purposes without their consent, depriving them of their autonomy, and ultimately – the lack of recognition of their intrinsic value. In a way, scientists have already committed an ethical decision to disregard these much more basic considerations by choosing to deploy animals in experiments in the first place. Yet, when the main rationale behind such a decision proves faulty, namely, that animal experimentation holds exceptional value for us, humans, then this should lead us to not only revise our welfare frameworks, but also our judgment of generally using animals in research. With this in mind, in the last paragraphs of my thesis I aim to briefly highlight these underlying ethical dilemmas that should be considered altogether within the epistemic context which critically examines the human benefit of animal experimentation.

Lack of Consent

From a consent perspective, there is a fundamental difference in how research is conducted with humans compared to animals. When human subjects are involved, an official and formal consent for participation is mandatory, regardless of the goals of the study (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979). The subjects are usually informed thoroughly about their role in the study, what exactly is expected from them and what they can expect, and all potential harms, hazardous aspects, and dangers are obligatory brought to their awareness. The process of formally adhering to a fully informed consent is safeguarding the individual’s personhood, autonomy, and fundamental rights (Baynham, 2022). It is important to mention that from an ethical research perspective, a person is considered as “vulnerable” if he/she has a diminished ability to fully safeguard their own interests within the

context of the study (Van Patter & Blattner, 2020). Thus, vulnerable participants could be more susceptible to intentional or unintentional exploitation and/or harm, which demands a higher level of ethical considerations, in order to protect one's dignity and prevent injustice (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979). For example, sometimes humans are also unable to provide consent due to various conditions such as severe illness or cognitive impairment. However, even in these situations the ethical and legal guidelines oblige that a consent must be acquired, usually through a family and/or legal individual, who represents the subject (Regulation (EU) No 536/2014). In fact, this is required even in situations that involve potentially life-saving interventions.

In stark contrast, animals subjected to experimentation are inherently unable to provide such consent, which raises significant moral concerns from an ethical perspective, as this already presupposes vulnerability. This problem only amplifies when animals are used in research where the procedure makes them suffer, or results in their death. But even if they could have provided such a consent – Would they? Who would subscribe for an experiment where the goal is that the individual develops endless self-harming behaviors? According to Rollin (2006), the inability of animals to provide consent, and therefore being forced into experiments, underlies a fundamental injustice, especially when distress and harm is caused. In reality the sole element that stands between researchers planning to use animal models in their studies and actually using them is an ethics committee that evaluates primarily to what extend the study's potential outcomes would benefit humans, or the so-called harm-benefit analysis (EU Directive, Article 38/63, 2010). Practically, this means that the bigger the assumed benefit for humans is, the bigger the tolerance to animals' suffering and harm would be. Although such regulations are enforced with good intentions, namely, to safeguard at least to some extend the welfare of animals, the very fact that a harm-benefit analysis exists in the first place is underlying a deeper ethical problem – animals used in experiments are deprived of their autonomy and reduced to mere tools for human interests.

Lack of Autonomy

If we consider that unlike humans, we do not (and partially cannot) require any consent from animals when it comes to deploying them in experiments, there is a particular consequence that emerges from an ethical standpoint, namely, that the used animals are deprived of their autonomy. Conceptually, being autonomous is typically associated with one's ability to self-govern, free from the control or manipulation of third parties (Christman, 2009). Animals that are subjected to

experimentation clearly do not possess any capacity to make a choice regarding their participation in a study. Even more than that, when the research involves causing harm, pain, or distress, it is obvious that they want to avoid and escape the experimental conditions which are making them suffer. In some experiments, however, the goal is precisely to put animals in a state of suffering. For instance, the very nature of the forced swim test, which I have described earlier in my thesis, puts rats and mice in conditions that lead them to a state of immobility, after desperately trying to escape the water container (Porsolt et al., 1977). Another example can be also derived from the first case study I focused on, where genetic modifications caused mice to self-harm to a point where open wounds and lesions occurred.

Although the cases in which animals are put under conditions of suffering/pain/harm are countless, there is a common ethical consequence which emerges, namely, the fact that humans dictated the fate of these other sentient beings by exerting power over them. From a moral perspective, this implies that animals do not have the right to act independently, and that their own interests and needs are being disregarded. In addition, by deciding over animals' lives, humans are imposing domination, which is a problematic act in any ethical discussion (Ingold, 1994). Essentially, the deprivation of choice applied to non-human animals showcases not only the negation of their capacity to self-govern themselves, but also demonstrates how humans are exercising their power superiority. Such actions are inherently contrasting any moral considerations and lead to important ethical questions regarding the responsibilities humans have towards sentient beings, especially towards those in their care. Yet, similarly to the ethical problem regarding the lack of consent, the problem of deprived autonomy again underscores an even deeper issue – that humans do not recognize the intrinsic value of animals.

Unrecognized Intrinsic Value

As we observed, the lack of consent in animal experimentation is inevitably connected to a deprivation of autonomy. But one might ask – what does the absence of autonomy imply, besides the obvious enforcement of power dynamics? Actually, on a deeper and more fundamental level, there is yet another layer to this ethical problem, namely, the disregard for the intrinsic value of animals. The concept of intrinsic value refers to the innate worth that a subject possesses on its own, irrespective of any external factors or purposes it may be related to. Essentially, intrinsic value is the recognition that a given entity has worth simply because it exists (O'Neill, 1992). Recognizing animals as ends in themselves affirms their inherent value and acknowledges that they exist for their

own sake (Regan, 1986). Contrary to that, if we are to determine their fate based mainly on weighing it against a potential human benefit, we are adopting the opposite view – that animals have merely an instrumental value. Having said that, in an experimental context where animal models are deployed, especially when harm and suffering are involved, the intrinsic value of animals is diminished to a mere resource which is useful only as a tool to pursue a scientific discovery. In that way, their inherent worth as sentient organisms is essentially being disregarded.

When it comes to humans who are subjected to experimentation, their choice to participate is not only respected, but we are given the freedom to opt out and cancel our participation at any given moment. Such an option underlies that even if someone had made the conscious choice of taking part in a research, this choice is by no means ultimate and can be changed based on unpleasant experiences or any other personal reason. Framing animals as instruments rather than recognizing their capacity to suffer and experience distress, is therefore diminishing not only the worth of their life as a whole, but also the importance of how this life is experienced. On this note, I would allow myself to make a claim here – Suppose we ask the very scientists that utilize animals in research the following question, with the option to answer with Yes or No – *"Do you think that the capacity of animals to experience joy, pain, or fear, are meaningful beyond our own (human) interest?"*. My intuition leads me to believe that a considerable part of the answers would be – Yes. But then, what would be the answer to the very question that inspired this ethical section, namely - What justifies subjecting animals to experiments that involve substantial pain, suffering, or death?

If there are scientists who indeed believe that animals do have an intrinsic value, then experimenting with such directly contradicts this belief. And if there are such that do not find the life of animals meaningful beyond our own benefit, then I would at least encourage them to carefully reflect on whether the presupposed epistemic benefit matches the final outcome. From my perspective, regardless of which side of the argument one picks, one should strive to contribute to and develop a more compassionate approach to scientific research in general.

To summarize it all, the discussion on the ethical dimensions of using animals in research – and also of using them for modeling purposes – posits several critical issues. To begin with, as consent cannot be obtained directly from the subjected animals, their use in experiments is solely left to the respective ethics committee to decide. The committees, however, in line with the laws on animal experimentation largely presuppose that animals can in principle be used. They rather advice on how to use them, weighing animal well-being and life against benefits for humans. Such a

weighing logic, however, neglects a range of more profound philosophical questions. Using animals, for example, usually means to not ask them for their consent and to negate their autonomy, which is especially problematic in situations where harm, pain, and/or suffering is intentionally inflicted on them. In addition, the fact that humans are planning, intending, and causing suffering to other sentient beings is deeply unjust from a moral perspective. The intrinsic value of the animals is unrecognized and they are merely seen as a resource (Regan, 1986). Considering all of the ethical aspects mentioned, as well as the critical examination of the epistemic justification from a methodological perspective, I believe that cognitive scientists should carefully review the methods that allow us to acquire knowledge and proactively shift towards more ethically responsible research practices.

Acknowledgments

I would like to express my heartfelt gratitude for all the support, encouragement, critical thought, and partnership, to my two exceptional supervisors – Ms. Zsofia Viranyi and Ms. Judith Benz-Schwarzburg, as well as to my dear and close friend Francesco Tonci Ottieri Della Ciaja, who inspired me to write this thesis with his endless determination to make our world a better one for all animals.

Bibliography

Abramowitz, J. S., & Jacoby, R. J. (2011). Obsessive-compulsive disorder in adults – Treatment of obsessive-compulsive disorder. Annual Review of Clinical Psychology, 7, 229-243. https://doi.org/10.1146/annurev-clinpsy-032210-104533

Akhtar A. The flaws and human harms of animal experimentation. Camb Q Healthc Ethics. 2015 Oct;24(4):407-19. doi: 10.1017/S0963180115000079. PMID: 26364776; PMCID: PMC4594046.

Andersen, M. L., & Winter, L. M. F. (2019). Animal models in biological and biomedical research: Experimental and ethical concerns. *Anais da Academia Brasileira de Ciências, 91*(suppl 1), e20170238. https://doi.org/10.1590/0001-3765201720170238

Avdeef, A. (2003). *Absorption and drug development: Solubility, permeability, and charge state*. John Wiley & Sons.

Bains, R. S., Cater, H. L., Sillito, R. R., Chartsias, A., Sneddon, D., Concas, D., Aspey, K., & Banks, G. (2016). Analysis of individual mouse activity in group housed animals of different inbred strains using a novel automated home cage analysis system. Frontiers in Behavioral Neuroscience, 10, 106. https://doi.org/10.3389/fnbeh.2016.00106

Banting, F. G., Best, C. H., Collip, J. B., Campbell, W. R., & Fletcher, A. A. (1922). Pancreatic extracts in the treatment of diabetes mellitus. *Canadian Medical Association Journal, 12*(3), 141-146.

Bard, K. A., Brent, L., Lester, B., Worobey, J., & Suomi, S. J. (2011). Neurobehavioural integrity of chimpanzee newborns: Comparisons across groups and across species reveal gene–environment interaction effects. Infant Child Development, 20(1), 47–93. *https://doi.org/10.1002/icd.686*

Baynham, K. (2022). Is it morally permissible for non-invasive, experimental cognitive research to be carried out with captive-bred chimpanzees? (Master's thesis). Messerli Research Institute, University of Veterinary Medicine Vienna.

Bekkali, S., Youssef, G. J., Donaldson, P. H., & Ismail, S. (2021). Is the putative mirror neuron system associated with empathy? A systematic review and meta-analysis. Neuropsychology Review, 31(1), 14–57.

Brewer, M. B. (2000). Research design and issues of validity. In H. T. Reis & C. M. Judd (Eds.), Handbook of research methods in social and personality psychology (pp. 3–16). Cambridge University Press.

Calvin, W. H. (1983). The throwing Madonna: Essays on the brain. New York, NY: MacGraw-Hill.

Carvalho, C., Peste, F., Marques, T. A., Knight, A., & Vicente, L. M. (2020). The contribution of rat studies to current knowledge of major depressive disorder: Results from citation analysis. Frontiers in Psychology, 11, 1486. doi: 10.3389/fpsyg.2020.01486

Capecchi, M. R. (2005). Gene targeting in mice: Functional analysis of the mammalian genome for the twenty-first century. Nature Reviews Genetics, 6(6), 507-512.

Chen, S.-K., Tvrdik, P., Peden, E., Cho, S., Wu, S., Spangrude, G., & Capecchi, M. R. (2010). Hematopoietic origin of pathological grooming in Hoxb8 mutant mice. Cell, 141(5), 775–785.

Christman, J. (2009). The Politics of Persons: Individual Autonomy and Socio-historical Selves. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511635571

Crabbe, J. C., Wahlsten, D., & Dudek, B. C. (1999). Genetics of mouse behavior: Interactions with laboratory environment. Science, 284(5420), 1670-1672.

Cryan, J. F., Markou, A., & Lucki, I. (2002). Assessing antidepressant activity in rodents: Recent developments and future needs. Trends in Pharmacological Sciences, 23(5), 238-245.

Daniele, A., Albanese, A., Contarino, M. F., Zinzi, P., Barbier, A., Gasparini, F., Romito, L. M. A., Bentivoglio, A. R., & Scerrati, M. (2003). Cognitive and behavioural effects of chronic stimulation of the subthalamic nucleus in patients with Parkinson's disease. Journal of Neurology, Neurosurgery & Psychiatry, 74(2), 175–182.

Davis, C., Kleinman, J. T., Newhart, M., Gingis, L., Pawlak, M., & Hillis, A. E. (2008). Speech and language functions that require a functioning Broca's area. Brain and Language, 105(1), 50-58. https://doi.org/10.1016/j.bandl.2008.01.012

DeGrazia, D. (1996). Taking animals seriously. Cambridge University Press. https://doi.org/10.1017/CBO9781139172967

DeGrazia, D. Animal Ethics Around the Turn of the Twenty-First Century. Journal of Agricultural and Environmental Ethics 11, 111–129 (1998). https://doi.org/10.1023/A:1009504617295

Dirnagl, U., & Endres, M. (2014). Found in translation. Stroke, 45, 1510–1518.

Di Pellegrino, G., Fadiga, L., Fogassi, L., & Gallese, V. (1992). Understanding motor events: A neurophysiological study. Experimental Brain Research, 91, 176–180. https://doi.org/10.1007/BF00230027

D'Hooge, R., & De Deyn, P. P. (2001). Applications of the Morris water maze in the study of the hippocampus. Behavioural Brain Research, 125(2-3), 167-180.

European Parliament and Council. (2014). Regulation (EU) No 536/2014 of the European Parliament and of the Council of 16 April 2014 on clinical trials on medicinal products for human use, and repealing Directive 2001/20/EC. Official Journal of the European Union, L158, 1-76. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32014R0536

European Parliament and Council. (2010). Directive 2010/63/EU of the European Parliament and of the Council of 22 September 2010 on the protection of animals used for scientific purposes. Official Journal of the European Union, L276, 33-79. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32010L0063

Gardner, R. A., & Gardner, B. T. (1969). Teaching sign language to the chimpanzee. Science, 165(3894), 664-672.

Garner, J. P., & Kutcher, G. S. (2013). Conceptual and methodological considerations for the use of nonhuman animals in psychological science: Historical precedents, contemporary concerns, and emerging paradigms. Psychological Bulletin, 139(2), 233-255.

Gallistel, C. R., & King, A. P. (2009). The role of animal models in cognitive science. In R. J. Sternberg & K. A. Sternberg (Eds.), Cognitive psychology (5th ed., pp. 495-510). Wadsworth.

Greenwald, A. G. (1976). Within-subjects designs: To use or not to use? Psychological Bulletin, 83(2), 314-320. *https://doi.org/10.1037/0033-2909.83.2.314*

Greek, R., & Menache, A. (2013). Systematic reviews of animal models: Methodology versus epistemology. *International Journal of Medical Sciences, 10*(3), 206–221.

Grimm, H., Olsson, I. A. S., & Sandøe, P. (2019). Harm–benefit analysis: What is the added value? A review of alternative strategies for weighing harms and benefits as part of the assessment of animal research. Laboratory Animals, 53(6), 17-27. DOI: *10.1177/0023677218783004*

Hallett, M. (2007). Transcranial magnetic stimulation: A primer. Neuron, 55(2), 187-199. https://doi.org/10.1016/j.neuron.2007.06.026

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? Behavioral and Brain Sciences, 33(2-3), 61-83. https://doi.org/10.1017/S0140525X0999152X

Herrmann, E., Call, J., Hernandez-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. Science, 317(5843), 1360-1366. https://doi.org/10.1126/science.1146282

Herndon, J. G., Moss, M. B., Rosene, D. L., & Killiany, R. J. (1997). Patterns of cognitive decline in aged rhesus monkeys. Neurobiology of Aging, 18(3), 239-248.

Hsiao, K., Chapman, P., Nilsen, S., Eckman, C., Harigaya, Y., Younkin, S., Yang, F., & Cole, G. (1996). Correlative memory deficits, Aβ elevation, and amyloid plaques in transgenic mice. Science, 274(5284), 99-102.

Iacoboni, M. (2009). Imitation, empathy, and mirror neurons. Annual Review of Psychology, 60, 653-670. https://doi.org/10.1146/annurev.psych.60.110707.163604

Ioannidis, J. P. A. (2005). Why most published research findings are false. PLOS Medicine, 2(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Ingold, T. (1994). Trust, Domination, and the Alternative History of Human-Animal Relations. In A. Manning & J. Serpell (Eds.), Animals and Human Society: Changing Perspectives (pp. 209-223). Routledge. Retrieved from *https://www.taylorfrancis.com/chapters/edit/10.4324/9780203421444-4/trust-domination-alternative-history-human-animal-relations-tim-ingold*

Jentsch, J. D., & Taylor, J. R. (2009). Impulsivity resulting from frontostriatal dysfunction in drug abuse: Implications for the control of behavior by reward-related stimuli. Psychopharmacology, 146(4), 373-390.

Kacelnik, A., & Bateson, M. (1996). Risky theories—the effects of variance on foraging decisions. American Zoologist, 36(4), 402–434. https://doi.org/10.1093/icb/36.4.402

Kimble, G. A. (1967). What's right and wrong with the Skinner box? The American Psychologist, 22(12), 887-896.

Kinter, L. B., DeHaven, R., & Johnson, D. K. (2021). A brief history of use of animals in biomedical research and perspective on non-animal alternatives. *ILAR Journal*, *62*(1–2), 7. Retrieved from https://academic.oup.com/ilarjournal/article-abstract/62/1-2/7/6309345

Kola, I., & Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? Nature Reviews Drug Discovery, 3, 711–716. https://doi.org/10.1038/nrd1470
Knight, A. (2016). "Systematic Reviews of Animal Experiments Demonstrate Poor Contributions to Human Healthcare " - speech at the international congress "Science instead of Animal Experiments" organized by Doctors Against Animal Experiments (DAAE) in Cologne, Germany

Knight, A. (2011). The costs and benefits of animal experiments. Palgrave Macmillan.

Le Dain, M.-A., Blanco, E., & Summers, J. D. (2013). Assessing design research quality: Investigating verification and validation criteria. In Proceedings of the International Conference on

Engineering Design (ICED13), 19–22 August 2013, Sungkyunkwan University, Seoul, Korea.

Leavens, D. A., Bard, K. A., & Hopkins, W. D. (2010). BIZARRE chimpanzees do not represent "the chimpanzee". Behavioral and Brain Sciences, 33(2-3), 100-101. doi: 10.1017/S0140525X10000166

Li, Q., Xiao, M., Liu, X., Zhao, Y., Yin, Y., Qiu, P., & Lu, F. (2023). Enhancing open clinical trials through blinded evaluations: An exploration with diabetic foot infections. Trials, 24(1), 150.

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. Nature, 453(7197), 869-878. https://doi.org/10.1038/nature06976

Mak, I. W., Evaniew, N., & Ghert, M. (2014). Lost in translation: Animal models and clinical trials in cancer treatment. American Journal of Translational Research, 6(2), 114–118. PMID: 24489990; PMCID: PMC3902221.

MacLean, E. L., Hare, B., Nunn, C. L., Addessi, E., Amici, F., Anderson, R. C., ... & Zhao, Y. (2014). The evolution of self-control. Proceedings of the National Academy of Sciences, 111(20), E2140-E2148.

Macy, A., Theo, J., Kaufmann, S., Ghazzaoui, R., Pawlowski, P., Fakhry, H., ... & IsHak, W. (2013). Quality of life in obsessive-compulsive disorder. CNS Spectrums, 18(1), 21-33. https://doi.org/10.1017/S1092852912000697

Makin, S., Cookson, M., & Lewis, P. (2016). The scientific ideas changing drug discovery. Nature News, 529(7586), S20-S21.

Maiers, W. (2022). Replication Crisis – Just Another Instance of the Replication of Crises in Psychology? Historical Retrospections and Theoretical-Psychological Assessments. Review of General Psychology, 26(2), 250-260. *https://doi.org/10.1177/10892680211033915*

Morris, R. G. M. (1984). Developments of a water-maze procedure for studying spatial learning in rats. Journal of Neuroscience Methods, 11(1), 47-57.

Nestler, E. J., & Hyman, S. E. (2010). Animal models of neuropsychiatric disorders. Nature Neuroscience, 13(10), 1161-1169.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). The Belmont Report: Ethical principles and guidelines for the protection of

human subjects of research. U.S. Department of Health and Human Services.

O'Neill, J. (1992). The varieties of intrinsic value. The Monist, 75(2), 119-137.

Paixão, R. L., & Schramm, F. R. (1999). Ethics and animal experimentation: What is debated Cadernos de Saúde Pública, 15(Suppl 1), 34-41.

Palmer, D., Dumont, J. R., Dexter, T. D., Prado, M. A. M., Finger, E., Bussey, T. J., & Saksida, L. M. (2021). Touchscreen cognitive testing: Cross-species translation and co-clinical trials in neurodegenerative and neuropsychiatric disease. Neurobiology of Learning and Memory, 182, 107443. https://doi.org/10.1016/j.nlm.2021.107443

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2009). Learning styles: Concepts and evidence. Psychological Science in the Public Interest, 9(3), 105–119.

Pavlov, I. P. (1927). Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex. Oxford University Press.

Petralia, R. S., & Wenthold, R. J. (1992). Light and electron immunocytochemical localization of AMPA-selective glutamate receptors in the rat brain. The Journal of Comparative Neurology, 318(3), 329-354. https://doi.org/10.1002/cne.903180309

Percie du Sert N, Ahluwalia A, Alam S, Avey MT, Baker M, et al. (2020) Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0. PLOS Biology 18(7): e3000411. https://doi.org/10.1371/journal.pbio.3000411

Poon, J. A., Thompson, J. C., Forbes, E. E., & Chaplin, T. M. (2018). Adolescents' reward-related neural activation: Links to thoughts of nonsuicidal self-injury. Suicide and Life-Threatening Behavior, 49(1), 76-89. https://doi.org/10.1111/sltb.12418

Pound, P., & Ritskes-Hoitinga, M. (2018). Is it possible to overcome issues of external validity in preclinical animal research? Why most animal models are bound to fail. Journal of Translational Medicine, 16(1), Article 304. https://doi.org/10.1186/s12967-018-1678-1

Rapp, A. M., Bergman, R. L., Piacentini, J., & McGuire, J. F. (2016). Evidence-based assessment of obsessive–compulsive disorder. Journal of Central Nervous System Disease, 8, 13–29. https://doi.org/10.4137/JCNSD.S38359

Regan, T. (1986). A case for animal rights. In M. W. Fox & L. D. Mickley (Eds.), Advances in animal welfare science 1986/87 (pp. 179-189). Washington, DC: The Humane Society of the United States.

Robbins, T. W., Vaghi, M. M., & Banca, P. (2019). Obsessive-compulsive disorder: Puzzles and prospects. Neuron, 102(1), 27–47. https://doi.org/10.1016/j.neuron.2019.01.046

Rosenthal, N., & Brown, S. (2007). The mouse ascending: Perspectives for human-disease models. Nature Cell Biology, 9(9), 993-999.

Rollin, B. E. (1995). The ethical status of animals. Springer.

Rouquette, A., & Falissard, B. (2011). Sample size requirements for the internal validation of psychiatric scales. Journal of Methods in Psychiatric Research, 20(4), 234-249.

Salkovskis, P. M. (1999). Understanding and treating obsessive-compulsive disorder. Behaviour Research and Therapy, 37(Suppl 1), S29-S52. PMID: 10402695.

Shmelkov, S. V., Hormigo, A., Jing, D., Proenca, C. C., Bath, K. G., Milde, T., Shmelkov, E., Kushner, J. S., Baljevic, M., Dincheva, I., Murphy, A. J., Valenzuela, D. M., Gale, N. W., Yancopoulos, G. D., Ninan, I., Lee, F. S., & Rafii, S. (2010). Slitrk5 deficiency impairs corticostriatal circuitry and leads to obsessive-compulsive–like behaviors in mice. Nature Medicine, 16(5), 598–602.

Sjoberg, E. A. (2017). Logical fallacies in animal model research. Behavioral and Brain Functions, 13(1), 3. https://doi.org/10.1186/s12993-017-0121-8

Slack, M. K., & Draugalis, J. R. (2001). Establishing the internal and external validity of experimental studies. *American Journal of Health-System Pharmacy, 58*(22), 2173-2181.

Skinner, B. F. (1953). Science and human behavior. The Macmillan Company.

Skinner, B. F. (1957). Verbal behavior. Appleton-Century-Crofts.

Skinner, B. F. (1974). About behaviorism. Knopf.

Stahnisch, F. W. (2008). Instrument transfer as knowledge transfer in neurophysiology: François Magendie’s (1783–1855) early attempts to measure cerebrospinal fluid pressure. Journal of the

History of the Neurosciences, 17(1), 72–99. https://doi.org/10.1080/09647040600913699

Supino, P. G. (2012). Fundamental issues in evaluating the impact of interventions: Sources and control of bias. In Principles of research methodology: A guide for clinical investigators (pp. 81-94). Springer.

Suddendorf, T., & Corballis, M. (2007). The evolution of foresight: What is mental time travel and is it unique to humans? Behavioral and Brain Sciences, 30, 299-351. https://doi.org/10.1017/S0140525X07001975

Sutherland, B. A., Minnerup, J., Balami, J. S., Arba, F., Buchan, A. M., & Kleinschnitz, C. (2012). Neuroprotection for ischaemic stroke: Translation from the bench to the bedside. International Journal of Stroke, 7(5), 407–418.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals (Series of Monograph Supplements, Vol. II, No. 4). The Macmillan Company.

Tong, C., Huang, G., Ashton, C., & Ying, Q. L. (2011). Generating gene knockout rats by homologous recombination in embryonic stem cells. Nature Protocols, 6, 827–844. https://doi.org/10.1038/nprot.2011.338

UNESCO. (2005). Universal Declaration on Bioethics and Human Rights. Paris: United Nations Educational, Scientific and Cultural Organization. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000146180.page=76

Van der Staay, F. J. (2006). Animal models of behavioral dysfunctions: Basic concepts and classifications, and an evaluation strategy. Behavioural Brain Research, 168(1), 1-16.

Van der Staay, F. J., Arndt, S. S., & Nordquist, R. E. (2009). Evaluation of animal models of neurobehavioral disorders. Behavioral and Brain Functions, 5, Article 11. https://doi.org/10.1186/1744-9081-5-11

Van Overwalle, F., & Baets, S. (2022). Individual differences in mirror neuron activity and empathy: A meta-analysis. Neuroscience & Biobehavioral Reviews, 141, 114-130.

Van Patter, L. E., & Blattner, C. (2020). Advancing ethical principles for non-invasive, respectful research with non-human animal participants. Society & Animals, 28(2), 171-190.

Walach, H., Falkenberg, T., Fønnebø, V., Lewith, G., & Jonas, W. B. (2006). Circular instead of hierarchical: Methodological principles for the evaluation of complex interventions. BMC Medical Research Methodology, 6(1), 29. https://doi.org/10.1186/1471-2288-6-29

Welch, J. M., Lu, J., Rodriguiz, R. M., Trotta, N. C., Peca, J., Ding, J.-D., Feliciano, C., Chen, M., Adams, J. P., Luo, J., Dudek, S. M., Weinberg, R. J., Calakos, N., Wetsel, W. C., & Feng, G. (2007). Cortico-striatal synaptic defects and OCD-like behaviours in Sapap3-mutant mice. Nature, 448, 894–898. https://doi.org/10.1038/nature06104

Wilke, C., Lahiff, N. J., Sabbi, K. H., Watts, D. P., Townsend, S. W., & Slocombe, K. E. (2022). Declarative referential gesturing in a wild chimpanzee (Pan troglodytes). Proceedings of the National Academy of Sciences, 119(47), e2206486119. https://doi.org/10.1073/pnas.2206486119

Wolfer, D. P., & Juraska, J. M. (2014). Measuring spatial memory in the Morris water maze. Journal of Visualized Experiments, (92), e52054.

Yankelevitch-Yahav, R., Franko, M., Huly, A., & Doron, R. (2015). The forced swim test as a model of depressive-like behavior. *Journal of Visualized Experiments*, *97*, e52587. https://doi.org/10.3791/52587