Robot Models of Mental Disorders

Matthew Lewis and Lola Cañamero
Embodied Emotion, Cognition and (Inter-)Action Lab
School of Computer Science
University of Hertfordshire, Hatfield, AL10 9AB, U.K.
Email: {M.Lewis4;L.Canamero}@herts.ac.uk

***Abstract*—Alongside technological tools to support wellbeing and treatment of mental disorders, models of these disorders can also be invaluable tools to understand, support and improve these conditions. Robots can provide ecologically valid models that take into account embodiment-, interaction-, and context-related elements. Focusing on Obsessive-Compulsive spectrum disorders, in this paper we discuss some of the potential contributions of robot models and relate them to other models used in psychology and psychiatry, particularly animal models. We also present some initial recommendations for their meaningful design and rigorous use.**

1. Introduction

Alongside technological tools to support wellbeing and treatment of mental disorders, models of these disorders (e.g., their underlying cognitive and affective mechanisms, etiology and symptoms) can also be invaluable tools to understand, support and improve these conditions. The new field of computational psychiatry [1] has started producing computer-based simulated models of some conditions. Complementing and going beyond these models, robots can provide more ecologically valid models that also take into account embodiment-, interaction-, and context-related elements. We argue that robot models of mental disorders have significant potential as a research tool to complement existing models and techniques. As we will see, their contributions can be at many levels: for example, they can help refine conceptual models and their implementations in animal models; they can improve our understanding of theoretical models and how they might complement each other by operationalizing them in ways that permit precise application and comparison of models; and they permit carrying out of studies that are not possible using other models for either methodological or ethical reasons.

In this paper, we outline the different types of models of mental disorders, both theoretical and concrete (particularly animal) models with a special focus on Obsessive-Compulsive Disorder (OCD). We also summarize their conditions of validity, main strengths and limitations, and discuss how robot models can build on and complement them, to make contributions to translational research. As work towards this goal, we propose an initial set of recommendations for the implementation of robot models, and an iterative design process based on a process from the animal model literature.

2. Models of Mental Disorders and OCD

In our research on robot models of mental disorders, we are currently focusing on Obsessive-Compulsive Disorder (OCD), and we will first provide a brief characterization of this condition as background to understand the discussion in this paper. Care must be taken with the word "model", as it can have different meanings in different scientific disciplines [2], and the rest if this section aims to clarify these different meanings. We distinguish two categories of model: conceptual (theoretical) models, and concrete implementations such as animal models and computational models.

2.1. Obsessive-Compulsive and Related Disorders

Obsessive-Compulsive Disorder (OCD) is a disabling mental health disorder characterized by obsessions (recurrent, invasive, often unpleasant thoughts) and/or compulsions (a strong urge to carry out certain repetitive or ritualized behaviors, such as hand washing or excessive checking). OCD is considered as part of the obsessive-compulsive (OC) spectrum, which also includes conditions such as trichotillomania (pathological hair pulling), body dismorphic disorder (BDD), and tic disorders such as Tourette's syndrome. In line with this, the Diagnostic and Statistical Manual of Mental Disorders, 5th ed. (DSM-V, 2013) introduced a category of Obsessive-Compulsive and Related Disorders (OCRD) which includes a number of these spectrum conditions.

The main treatments of OCD are psychological interventions—such as exposure and response prevention (ERP)—and medication—usually SSRIs (selective serotonin reuptake inhibitors which, as the name suggests, reduce the rate of reabsorption of the neurotransmitter serotonin). However, these existing treatments fail to help a significant portion of patients [3].

The superficially related condition of obsessive-compulsive personality disorder (OCPD) is characterized by excessive perfectionism, and desire for "orderliness" (e.g., a needless desire for symmetry) and control. The main difference between OCD and OCPD is that OCPD is part of the person's personality and therefore perceived by them as normal, rather than unwanted. However, whether OCPD should be considered within the OC spectrum is an open question [4].

2.2. Conceptual models

By a "conceptual model of a mental disorder" we mean a theoretical construct that links underlying causes (etiology), either proposed or observed, with observed symptoms and correlates. A conceptual model should serve as a framework for understanding, and should have explanatory and predictive power with respect to the condition being modeled. Conceptual models may be based on psychological, neuroscientific or other frameworks for understanding cognition and behavior. A conceptual model may include currently theorized elements of the mental disorder that elaborate a hypothesized mechanism. The related term "causal model" is widely used and may be considered as meaning a conceptual model that describes causal relations between the elements of a system. In practice, many conceptual models are therefore causal models.

As examples of conceptual models we can mention the five cognitive-behavioral models for OCD discussed by Shafran [5]. These variously propose as the key mechanism: (i) a faulty appraisal of normal intrusive thoughts due to an inflated sense of responsibility, (ii) an inflated appraisal of the significance of normal intrusive thoughts, (iii) an excessive emphasis on control of one's own thoughts, (iv) faulty evaluation of the likelihood and consequences of danger, (v) a self-perpetuating mechanism in which checking behavior fails to provide satisfactory certainty, but increases sense of responsibility and thereby elevates the probability of harm.

It is important to note that there is not one "true" model that we are seeking; different models may be compatible with each other, but have different emphases or different levels of abstraction, and it would be beneficial to understand their complementarities. It may also be that a mental health condition currently viewed as a single entity has multiple mechanisms or causes and is therefore better viewed as multiple conditions, each with different models. We claim that robot models can contribute to this with a synthetic approach that complements and supports analysis.

2.3. Concrete models

2.3.1. Animal models. Van der Staay gives the following definition of an animal model [6]:

> "An animal model with biological and/or clinical relevance in the behavioral neurosciences is a living organism used to study brain–behavior relations under controlled conditions, with the final goal to gain insight into, and to enable predictions about, these relations in humans and/or a species other than the one studied, or in the same species under conditions different from those under which the study was performed."

Animal models may be induced by genetic techniques (e.g., gene knockout in experimental animals), by the use of drugs to induce symptoms similar to those of the disease, or by environmental manipulation (e.g., by introducing particular stressors or by using behaviorist approaches). Alternatively, they may be naturally occurring. As we shall see in section 3.2, an animal model will often have an underlying conceptual model associated with it.

Animal models have the advantages that they model complete systems (organism and environment), and they use a real animal, hence a real nervous system. However, there are limits to how closely a non-human animal can be used to model human mental disorders. In addition, symptoms that can be inferred from animal models are largely those based on observation of behavior, with the internal experience that is important in descriptions of human mental disorders either inferred from behavior or from neurological examination of the animal (see section 3.4). Another important disadvantage of animal models is the ethical issues associated with animal experimentation. These issues become more problematic as the model animal is evolutionarily closer to humans—experimentation on, and housing of, primates is more ethically problematic than on mice.

2.3.2. Computational Models. Computational models are realizations, or partial realizations, of theoretical models in computers. The emerging field of computational psychiatry includes within its scope the development of computational models of psychiatric disorders [1].

With respect to animal models, these models have the advantage that, by their nature, they are highly specified and so any results should be replicable (or, preferably, reproducible [7]). Also, unlike animal models, the whole (modeled) system can be analyzed in detail, with the possibility that some internal processes can be related to human internal experience. However, due to the complexity of implementing such a model, for practical reasons they are typically only partial implementations (e.g., of a neurological subsystem) or they work at a relatively high level of abstraction (e.g., with brain areas, rather than individual neurons). In addition, they do not necessarily include any behavioral element, a true closed-loop (physical and social) interaction with the environment, or the effects of contextual and environmental elements.

2.3.3. Robot Models. Robot models of mental disorders, like computational models, include an embedded realization of a conceptual model. However, in this case the model is implemented in an embodied, interacting robot and its environment, which introduce new elements that purely computational models lack.

Embodiment is an important aspect of many mental disorders (e.g., distorted perception of the body, distorted perception of body ownership), as well as of some therapeutic approaches (e.g., exercise, art therapy). The embodied element of cognitive-affective autonomous robots permits a realistic and ecologically valid modeling of such phenomena, and the potential use of robots in "embodied" therapeutic practices. Embodied robot models can naturally take account of the fact that action modifies perception[1], both of the environment and of the own body, and this in turn affects action and the perception-action loop (which includes cognition at different levels). We argue that a more systematic understanding of dysfunctions in the perception-action loop might be key towards understanding mental disorders.

1. In some versions of the sensorimotor approach to cognition such as [8], action is a necessary component to account for the quality of sensory experience.

In addition, robots are situated in a physical (and often social) environment that can be very similar to a human environment in ways that are relevant to the study of cognition and their dysfunctions under specific circumstances and contexts. The use of embodied robots as models of mental disorders therefore also means that behavior can be modeled, observed and manipulated systematically in context, going beyond computational psychiatry approaches that model brain (dys-)functionality in neural networks or other machine learning algorithms. The fact that, due to their embodiment, robots can have real (as opposed to simulated) interactions with the (human-inhabited) environment is also crucial as it permits to take into account external triggers related to mental disorders. The interaction with the environment is important in situations where according to the conceptual model, the initial cause, or alternatively the trigger for symptoms, is related to environmental interactions.

Robot models can complement and help to understand animal models and their relevance to the understanding and treatment of human mental disorders in a number of important ways, as we will see in section 4.

3. Animal Models of Mental Disorders

In order to evaluate the potential of and to consider how to make the best use of robot models of mental disorders, we look at animal models in general and in their application in OC spectrum disorders.

3.1. Development of Animal Models

Van der Staay describes the development of animal models as an iterative process [6] that starts with selection of the phenotype to be modeled (e.g., a behavior, or an internal "endophenotype" such as a hormone imbalance), and continues until either no refinement of the animal model is required following testing (accept the model), or the model is considered inadequate and development is halted. In practice, the animal model may be continuously in development throughout its lifetime since it will never fully represent the target condition. Van der Staay's presentation of the development process does not explicitly mention an underlying theoretical model for the animal model. The fact that van der Staay's evaluation criteria (see section 3.2) allow an animal model to lack construct validity if it has face validity, seems to agree with the observation that the aspects to be modeled can include elements from the behavior or "phenotype" (leading to face validity) and the underlying aspects or "endophenotype" (leading to concept validity).

3.2. Evaluation

Animal models for human mental disorders are generally evaluated according to four criteria [6], [9], [10]:

- *Predictive validity*: Performance in the animal model predicts performance in the condition being modeled. For example, the animal model should successfully discriminate between effective and ineffective treatments.
- *Face validity*: Phenomenological similarity between the animal model and the condition. For example, the symptoms of the condition are observed in the animal.
- *Construct validity*: The same physiological, psychological or conceptual constructs are applicable in the animal model and the condition (homology). For example, brain lesions in the animal are made where brain damage is observed in humans.
- *Reliability*: The outputs of the animal model are robust and reproducible. For example, they can be reproduced between laboratories.

Note that the importance and interpretation of the three validity criteria may vary according to the goals of the animal model. For example, for clinical purposes, predictive validity would typically be of the highest importance as it leads to potential treatments.

Construct validity can be interpreted in different ways, but one aspect is as an assessment of the extent to which the animal model embeds a conceptual model of the condition. In this way an animal model with high construct validity can be used as evidence for (or against) a conceptual model if it shows high (or low) predictive and face validity. However, construct validity is also be seen where there are observations of endophenotypical correlation (see next section).

3.3. Animal Models of OC Spectrum Disorders

Camilla d'Angelo *et al.* examined twenty-nine animal models of OC spectrum disorders according to the above criteria [11]. In this section we summarize their main findings.

Many models achieved good face validity, with a common negative being that the animal behaviors were relevant to other disorders as well (lack of specificity). However, the behaviors observed were limited to the compulsive (behavioral) side of the OC spectrum, neglecting the (cognitive) obsessions. This is a general limitation of animal models with respect to mental health conditions, and therefore we discuss it in a separate section (see section 3.4). These problems have direct implications for the elaboration of robot models of OCD and more generally of mental disorders. From the point of view of researchers creating robot models, the issues related to the lack of specificity need to be considered at the design stage: if our model shows face validity for a targeted condition (e.g., OCD), what alternative diagnoses (e.g., OCD versus OCPD versus addiction) might be made based on those symptoms (e.g., repetitive behavior)? The characteristics of the model can be shaped so that more information about the internal endophenotype is available. For example, by explicitly modeling a particular brain region or receptor within the system, so that data can be collected from it.

In the above-mentioned study, construct validity was largely limited to endophenotypical correlations (for example, changes in hormone levels or the involvement of brain regions associated with the OC spectrum) rather than the embedding of a conceptual model. This may be due to a limitation in experimenters' ability to manipulate biology and thus to build a specific model. However, it is worth noting that in several cases the methods

used in creating the model were linked to OC spectrum endophenotypes (e.g., drugs were chosen with observed links to OCD) so this could be considered as an implicit underlying model.

Predictive validity was variable, often unknown, and largely assessed in terms of drug response (one case went beyond this since the model suggested the possible involvement of the immune system, with implications for human research). This is illustrative of a limitation of animal models: many psychological interventions are not easily testable, or only testable in limited ways. For example, exposure and response prevention (ERP) therapy used to treat OCD involves buy-in from the patient: they must be willing to undergo the exposure and resist responding (with support from the psychologist or friends), and this aspect is not possible to reproduce with animals.

In section 4.4 we discuss how robot models of mental disorders can be evaluated in terms of the same four validity criteria used for animal models, and how they complement animal models and permit to overcome the limitations outlined here. The design recommendations presented in section 5.1 also take into account these limitations and propose ways to overcome them when designing and using robot models of mental disorders in general, including the OC family.

3.4. Face Validity in Mental Disorders

As mentioned in the previous section when discussing animal models of the OC family, animal models of mental disorders tend to show good face validity, but present the problem of lack of specificity (i.e., the animal behaviors associated with a specific disorder are relevant to other disorders as well). Here we examine this problem in further detail.

The utility of face validity for animal models can be viewed as problematic when we look at how psychological symptoms are considered in the clinic. The classic text by Fish [12] groups symptoms into the following chapters:

- Disorders of Perception (e.g., hallucinations)
- Disorders of Thought (e.g., obsessive thoughts, delusions) and Speech (e.g., mutism, word deafness)
- Disorders of Memory (e.g., amnesias, déjà vu)
- Disorders of Emotion (e.g., abnormal emotional reactions)
- Disorders of the Experience of the Self (e.g., loss of sense of boundary with the environment, sense of alienation from one's own actions)
- Disorders of Consciousness (e.g., confusion, dream-like state)
- Motor Disorders (e.g., tics, stupor, sense of alienation from one's own actions)

This list highlights the difficulty in using these symptoms with animal models. While motor disorders such as tics or stupor can be read in animals, and amnesias can be probed with learning tasks, other symptoms are difficult to imagine being recognizable in animals. How could we recognize a sense of alienation from one's own actions in a mouse, or even a primate? This echoes the observation in section 3.3 that face validity in OCD spectrum animal models only addressed the compulsions, not the obsessions.

Taking behavior as a proxy for mental state must be done with caution even in apparently simple cases: mutism (equivalently, lack of vocalizations in an animal) could be caused by many different (physical and) mental states, that would correspond to quite different diagnoses in humans. By way of example, we recall the difference between OCD and OCPD in humans (section 2.1) in which similar behavioral patterns are not viewed as being on the same spectrum, in part due to the way they are perceived internally by the patient.

Neurologically invasive procedures may offer some insight into these internal aspects of a condition by way of neurological correlates of mental states or perceptions. However, they are less ethically acceptable than non-invasive techniques, especially in animals that are evolutionarily closer to humans. This is an area where robot models may offer benefit, since the whole "brain" is available for realtime tracking with little practical difficulty and without similar ethical issues.

4. Robot Models

4.1. Examples of Robot Models

Work in biologically-inspired cognitive robotics can help to elucidate cognitive dysfunctions. Embodied cognition and its development has been investigated and modeled in areas such as cognitive and developmental robotics for decades now, and these fields are sufficiently mature to undertake a well grounded study of its dysfunctions.

To date, very few robot models specifically of mental disorders have been developed, with work by Yamashita and Tani [13] being one of the rare examples. They modeled schizophrenia by first training a robot controlled by a neural network to move an object in specific ways. After training the robot, they followed a model of schizophrenia as a failure in top-down (prediction) and bottom-up (perception) interactions, by adding noise to weights connecting levels of the hierarchical network, thus introducing errors. This led to behaviors in robot such as "catalepsy" (freezing in place) and stereotypical (repetitive) behavior.

Other work in cognitive-affective and developmental robotics has also modeled more implicitly, or can give rise to the emergence of mental disorders as dysfunction of different cognitive-affective cognitive capabilities in robots, such as schizophrenia stemming from malfunctions of dopaminergic and serotonergic modulation in neuromodulated robots [14], addiction, impulsivity and compulsive behavior as dysfunctions of pleasure neuro-/hormonal modulation of incentive salience in decision making [15], or the development of anxious behavioral phenotypes and dysfunctions of attachment to a human caregiver as a result of poor or negative early sensorimotor and interaction experiences [16].

4.2. Advantages of Robot Models

One significant advantage of computational models, including robot models, over animal models is that it allows precise operationalization and explicit implementation of an underlying theoretical model. While models

may be implemented in animals, experimenters may not always have enough control over the biology to implement it as precisely as desired (see section 3.3).

In relation to our ability to replicate experiments, while there are variations between robots, in general these are more controlled than in animal models where differences in strain, gender, age, development/rearing conditions and health may lead to differences in experimental outcomes affecting reliability. For replicability, source code can be provided (with robot schematics, if necessary), and care should also be taken to note hardware and firmware revision numbers and configuration of the robots. The robots' "health" may have an effect, for example due to physical wear-and-tear on components. If possible, unique identifiers of robots should be recorded with the data in order that they can be identified and re-examined in future—a possibility not available with animals. Arguably more significant than replicability is reproducibility [7], and for this the description of the software is needed, along with its key theoretical underpinnings (the conceptual model) so that it can be re-implemented, potentially in a different robot platform.

Possible environmental confounding factors can also be better controlled with robots than with animals. For example, experimenter smell was unexpectedly found to affect mouse and rat stress responses [17], which may have affected results. Such effects can be limited in robots since we have more knowledge of, and control over, their sensory systems. However, care still needs to be taken, e.g., to control for variations in temperature, or lighting when using robots that use infra-red or light-based sensors.

Compared to animal models, robot models offer some flexibility in terms of embodiment, which can even be custom designed. Animal models are limited to existing organisms, with ethical issues limiting the use of some animals more than others. An example of where this may be of use is when the model agent is required to manipulate objects which may be difficult for animals, but for which a custom tool can be added to a robot.

Robot models also allow us to test hypothetical pharmaceutical interventions. For example, they may test simulated drugs that have a single targeted effect on one element, when no such chemical has yet been found, or when known chemicals have undesirable properties such as side effects or not crossing the blood-brain barrier. Such simulations could be used in helping to select targets for drug designers.

4.3. Disadvantages of Robot Models

Although robot models form a complete system (robot/environment), unlike in animal models we do not get a complete biological system with its associated complexities. For example, in examining the effects of a drug, factors such as absorption, the effect of the blood-brain barrier, and breakdown and excretion are automatically included when using animal models—in addition to possible unanticipated side effects. This is of crucial importance in predicting clinical effects.

Compared to purely computational models, many robot models will need to be simplified or use a higher level of abstraction due to the often reduced computational power, and the additional computational overheads related to being a complete agent entertaining complex sensorimotor interactions with the world. In addition, there are practical aspects of running robot experiments—e.g., keeping batteries charged, working in realtime (sometimes slower than realtime), physical breakdowns, managing the environment—that prevent large numbers of experiments being done for statistical validity.

There may also be limitations in the ability of current robots to perform many complex tasks, such as the fine manipulation that can be observed in OCD, for example, when objects are placed precisely at the "correct" position to make them symmetrical. This mirrors problems in animal models, which are limited in what human-like activities they can perform, although in many cases they can execute complex tasks, that we cannot yet implement in robots (we may not even have a good understanding of how the animals are able to do the task).

4.4. Evaluation of Robot Models

Robot models of mental disorders can be evaluated and validated along the same four criteria discussed for animal models, complementing their contributions to cross-disciplinary and translational research:

- *Face validity*: phenomenological similarity between a robot model of a condition, and the condition in humans (or in other animals) is a natural consequence of robot models, particularly in terms of behavior and interaction. As discussed previously, generating behavior and embodied interaction is the natural thing to do with robots, and an aspect where robots can best complement and go beyond computational models. Given a theoretical model of a cognitive-affective capability and hypotheses about its dysfunctions in humans, the operationalization of such a model in an embodied interacting robot will naturally seek to replicate behavioral phenotypes predicted and observed in humans in similar contexts. The face validity of robot models can thus be very high.
- *Construct validity*: the construct validity of animal models can be considered an issue or regarded as unproblematic depending on the theoretical perspective adopted. For example, approaches grounded in behavioristic psychology will more naturally accept that results obtained for one species will apply to other species since the underlying cognitive mechanisms (e.g., conditioned learning) are postulated to be similar. On the other hand, the construct validity of animal models will be more critically questioned by approaches that give much importance to species-specific features and differences, such as models grounded in ethology. The construct validity of robot models can be questioned on different grounds. For example, the fact that robots and biological systems are made of different matter; or that the models and algorithms implemented in robots are simplifications of biological constructs. However, what critics consider as weaknesses of these models can also be considered as strengths. The fact that robot models

are simplifications permits to capture key selected structural, functional, or dynamics elements for a focused, rigorous investigation. Different models of the same phenomenon that approach it at different levels of granularity from different theoretical perspectives can also be implemented, tested, and compared, permitting to bridge gaps across levels and conceptual perspectives, which is a crucial issue in cross-disciplinary and translational research. Interdisciplinary work to elaborate robot models is of course a must to guarantee the relevance and construct validity of the model.

- *Predictive validity*: given a robot model of a cognitive-affective capability or a condition with good construct and face validity, predictive validity stems naturally. In fact, predictive validity is one of the key methods used in biologically-inspired robotics to assess models, both qualitatively and quantitatively. In addition to assessing the accuracy of the prediction, robot models can allow us to understand the mechanisms and processes underlying the predicted behavior and how both relate, since it is possible to carry out detailed and rigorous quantitative measurements and analysis of the internal architecture of the robot and its behavior as a function of internal and external influences (e.g., stimuli) and their interaction dynamics (see e.g., [15], [16]).
- *Reliability*: good robot models should be very robust and highly reproducible. Since robots operate in the real world and due to embodiment-related features such as noise in sensors and actuators, it is never possible to reproduce exactly the same experimental conditions (as it would be possible with computer simulations). However, well designed controlled experiments permit the reproduction of similar key relevant conditions (controlled variables in both the environment and the robot) with enough accuracy; experiments can be repeated many times, and accurate statistical/mathematical analysis guarantees the significance and reliability of the results. Compared to animal models, robot models can be considerably more reliable in terms of controllability, replicability and robustness (see section 4.2).

5. Building Robot Models

In van der Staay's description, development of an animal model starts with either the selection of (endo)phenotypes [6] or a preliminary hypothesis stage, followed by the selection of (endo)phenotypes [18]. In designing a robot model, we caution that it is very easy to create behaviors (phenotypes). Trivially, specific behaviors can simply be programmed if the designer wishes, but aberrant behaviors can also be created by poorly chosen parameters, faulty perception, or programming errors. Therefore, we will not take the phenotype as a starting point for robot models.

With a robot, unlike in animal models (see construct validity in section 3.3), we are free to take as a starting point a conceptual model, that may be implemented in the robot either with refinement (for example, if the conceptual model is insufficiently specified) or simplification. This gives a level of conceptual validity as part of the design process. However, it is important to note that this validity is with the conceptual model, not necessarily with the condition being modeled, indeed the robot model's success in generating the phenotype of the condition can serve as a test of the conceptual model.

5.1. Recommendations

In order to ensure that the potential of robot models for mental disorders in met, we propose the following initial list of recommendations for using robots as models for mental disorders:

1) Start design from a theoretical model (section 5).
2) The model must be capable of creating "good" phenotypes ("correct behavior") as well as symptomatic phenotypes (section 5).
3) Design and data collection should take into account endophenotypes: known endophenotypes should be accessible, and unobserved endophenotypes can be anticipated, especially those predicted by the conceptual model. In addition, in order to advance beyond animal models, design and data collection should consider "internal" symptoms (sections 3.3, 3.4).
4) Design should consider superficially similar disorders and consider how they might be distinguished. Do the chosen phenotypes have sufficient specificity? (sections 3.3, 3.4).
5) To aid reliability, the environment should be controlled, particularly with respect to elements that are likely to affect the functioning of the robot (sections 4.2, 4.4).
6) To aid replicability (as opposed to reproducibility) source code should be stored, along with details of the robots used and their configuration (section 4.2).
7) For translational research, predictive validity must be considered (section 3.2). In particular, the design should anticipate the integration of possible treatments. These could be environmental manipulations (taking advantage of the embodied aspect of the robot), simulated pharmaceuticals (see section 4.2), or simulated psychological interventions.
8) Models should differentiate between cure (targeting what maintains a disorder) and prevention (targeting the initial cause of a disorder), bearing in mind that patients may only present themselves after suffering for years with a condition.

5.2. Design Process

Based on the above discussion and recommendations, we propose the following iterative process for designing a robot model of a mental disorder (figure 1) based on van der Staay's process for animal models, and illustrated with some examples from OCD:

1) Select or create a suitable conceptual model of a mental disorder (or a small set of models that can

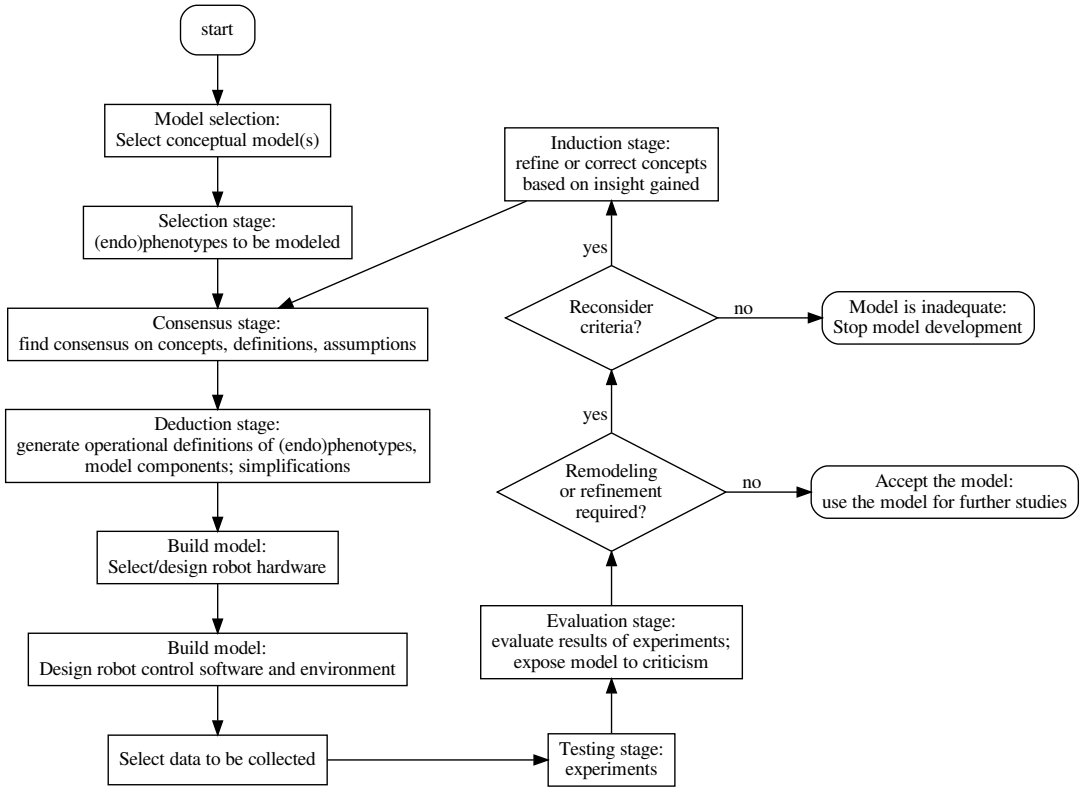

Figure 1. Flowchart for the proposed iterative process for designing a robot model of a mental disorder. This closely follows the process described in [6], [18]. Each stage is elaborated in section 5.1.

be implemented and compared). This will form the basis of the robot model, and in some cases ideas about the complementarity of models will allow multiple conceptual models to be combined or compared.

For OCD, compulsions can be understood as maladjusted reversal learning, as a problem related to signal attenuation, or as related to anxiety [19]. The choice of one or another will lead to the possibility of modeling different aspects and symptoms of OCD.

2) Select (endo)phenotypes of interest. This includes both behavioral phenotypes, internal cognitive phenotypes, and biological endophenotypes—see recommendation 3.

   For OCD, behaviors (phenotypes) could include repeatedly checking the same area of the environment (a "home" or "nest"). Animal models might usefully serve as inspiration for concrete behaviors that the robot can exhibit, as animal models provide simplified versions of human behavior. Going beyond these behavioral phenotypes, in a robot, in which we can examine the internal state, we can also have the opportunity to consider cognitive phenotypes such as obsessions, which are not easily observable in animal models. For example, this could be the dysfunctionally frequent triggering of a sensorimotor or cognitive representation associated with a compulsive behavior.

3) Consensus stage: find consensus on and refine concepts, criteria, definitions, assumptions. Care should be taken that the concepts agree with those in the psychological literature, and that there is sufficient specificity for the condition (see section 3.3 for the limitation in animal models, and recommendation 4).

4) Deduction stage: create operational definitions of the (endo)phenotypes selected, concepts. Some simplification may be required here, in order to take into account the limitations of the robot platform.

   For OCD, the checking behavior may be conceptually simplified as a repeated visiting behavior. How an internal aspect such as obsessions can be defined may depend on the control software to be used. For example, if the robot is controlled by a neural network, obsessions may correspond to repeated activation of a particular area of the network.

5) Model building: select/create robot hardware (it must at least be capable of the behaviors). Standard robot platforms can be used, or increasingly custom platforms can be designed and shared through standard controllers and 3D-printing.

   For OCD, if the phenotype of interest is visiting (checking) a particular area, then a simple mobile robot with appropriate sensors and simple behaviors to navigate its environment is sufficient (versus, for example, a sophisticated humanoid robot with complex social communication skills).

6) Model building: design robot control software (based on the refined theoretical models from step 3) and environmental features (which should

match the capabilities of the robot, as in the example in the previous step).

7) Select data to be collected. This should take into account recommendation 3: ideally, the literature should be searched for known and hypothesized (endo)phenotypes and if an analogue exists in the model, data can be collected relating to it.
   For OCD, in addition to data to quantify the (endo)phenotypes of interest, data can be collected related to alternative models. For example, review of the models for OCD described by Shafran and outlined in section 2.2 may lead to collecting data relating to the robot's internal appraisal mechanisms, evaluation of danger, or levels of certainty about relevant aspects of the environment.
8) Robot model testing: run experiments in the environment, collect and analyze data. The testing should also ensure that the model can produce non-pathological (endo)phenotypes, such as normal (non-compulsive) checking behavior in the case of OCD.
9) Model evaluation. The robot model can be evaluated with respect to the predictions of the conceptual model (both the underlying one and alternatives), animal models, purely computational models, and the mental health condition as it occurs in humans. In addition, predictions of the robot model can be tested in animal models or experiments with human patients. See section 4.4.
10) Refine or correct concepts and return to step 2. This may include modifying the conceptual model that underlies the robot model, and should involve collaboration with experts in the associated mental health condition, since it is the health condition that is the ultimate target of the model.

6. Conclusion

Robot models of mental disorders have the potential to add to our knowledge of these mental health conditions, to make clinical contributions in translational work, and to complement animal models and purely computation models, going beyond them in some aspects. However, in order to achieve this potential, we need to learn from the extensive literature on animal models, to take seriously its insights, and to note where its areas of weakness lie, in order to best complement it. As work towards this goal, we have proposed an initial set of recommendations, and an iterative design process based on a process from the animal model literature.

Acknowledgments

The first author is supported by an Early Career Research Fellowship on Robots as Embodied Models of Mental Disorders from the University of Hertfordshire. The authors would like to thank our collaborators, in particular, Prof. Naomi Fineberg (Hertfordshire Partnership NHS Foundation Trust, University of Hertfordshire and University of Cambridge) and Dr David Wellsted (NIHR East of England Research Design Service and University of Hertfordshire).

References

[1] Q. J. M. Huys, T. V. Maia, and M. J. Frank, "Computational psychiatry as a bridge from neuroscience to clinical applications," *Nature Neuroscience*, vol. 19, no. 3, pp. 404–413, 2016.

[2] R. Frigg and S. Hartmann, "Models in science," in *The Stanford Encyclopedia of Philosophy*, Spring 2017 ed., E. N. Zalta, Ed. [Online]. Available: https://plato.stanford.edu/archives/spr2017/entries/models-science/

[3] N. A. Fineberg and T. M. Gale, "Evidence-based pharmacotherapy of obsessive-compulsive disorder," *The International Journal of Neuropsychopharmacology*, vol. 8, no. 1, pp. 107–129, 2005.

[4] N. A. Fineberg, S. Reghunandanan, S. Kolli, and M. Atmaca, "Obsessive-compulsive (anankastic) personality disorder: Toward the ICD-11 classification," *Revista Brasileira de Psiquiatria*, vol. 36, pp. 40–50, 2014.

[5] R. Shafran, "Cognitive-behavioral models of OCD," in *Concepts and Controversies in Obsessive-Compulsive Disorder*, J. S. Abramowitz and A. C. Houts, Eds. Boston, MA: Springer, 2005, pp. 229–260.

[6] F. J. van der Staay, "Animal models of behavioral dysfunctions: Basic concepts and classifications, and an evaluation strategy," *Brain Research Reviews*, vol. 52, no. 1, pp. 131–159, 2006.

[7] C. Drummond, "Replicability is not reproducibility: Nor is it good science," June 2009. [Online]. Available: http://cogprints.org/7691/

[8] J. K. O'Regan, *Why Red Doesn't Sound Like a Bell: Understanding the Feel of Consciousness*. Oxford University Press, 2011.

[9] P. Willner, "Validation criteria for animal models of human mental disorders: Learned helplessness as a paradigm case," *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 10, no. 6, pp. 677–690, 1986.

[10] M. A. Geyer and A. Marcou, "The role of preclinical models in the development of psychotropic drugs," in *Neuropsychopharmacology: The Fifth Generation of Progress*, K. L. Davis, J. T. Coyle, and C. Nemeroff, Eds. Philadelphia, PA: Lippincott, Williams & Wilkins, 2002, pp. 445–455.

[11] L.-S. Camilla d'Angelo, D. M. Eagle, J. E. Grant, N. A. Fineberg, T. W. Robbins, and S. R. Chamberlain, "Animal models of obsessive-compulsive spectrum disorders," *CNS Spectrums*, vol. 19, no. 1, pp. 28–49, 2014.

[12] F. Fish, *Clinical Psychopathology: Signs and Symptoms in Psychiatry*, 1st ed. Bristol, UK: John Wright & Sons Ltd, 1967.

[13] Y. Yamashita and J. Tani, "Spontaneous prediction error generation in schizophrenia," *PLoS ONE*, vol. 7, no. 5, pp. 1–8, 05 2012.

[14] J. L. Krichmar, "A neurorobotic platform to test the influence of neuromodulatory signaling on anxious and curious behavior," *Frontiers in neurorobotics*, vol. 7, no. 1, 2013.

[15] M. Lewis and L. Cañamero, "Hedonic quality or reward? A study of basic pleasure in homeostasis and decision making of a motivated autonomous robot," *Adaptive Behavior*, vol. 24, pp. 267–291, 2016.

[16] J. Lones, M. Lewis, and L. Cañamero, "From sensorimotor experiences to cognitive development: Investigating the influence of experiential diversity on the development of an epigenetic robot," *Frontiers in Robotics and AI*, vol. 3, 2016.

[17] R. E. Sorge, L. J. Martin, K. A. Isbester, S. G. Sotocinal, S. Rosen, A. H. Tuttle, J. S. Wieskopf, E. L. Acland, A. Dokova, B. Kadoura, P. Leger, J. C. S. Mapplebeck, M. McPhail, A. Delaney, G. Wigerblad, A. P. Schumann, T. Quinn, J. Frasnelli, C. I. Svensson, W. F. Sternberg, and J. S. Mogil, "Olfactory exposure to males, including men, causes stress and related analgesia in rodents," *Nature Methods*, vol. 11, no. 6, pp. 629–632, 2014.

[18] F. J. van der Staay, S. S. Arndt, and R. E. Nordquist, "Evaluation of animal models of neurobehavioral disorders," *Behavioral and Brain Functions*, vol. 5, no. 11, 2009.

[19] N. A. Fineberg, S. R. Chamberlain, E. Hollander, V. Boulougouris, and T. W. Robbins, "Translational approaches to obsessive-compulsive disorder: From animal models to clinical treatment," *British Journal of Pharmacology*, vol. 164, no. 4, pp. 1044–1061, 2011.