cpsy

A Robot Model of OC-Spectrum Disorders: Design Framework, Implementation and First Experiments

Matthew Lewis[1], **Naomi Fineberg**[2,3] **and Lola Cañamero**[1]

[1]Embodied Emotion, Cognition and (Inter-)Action Lab, School of Computer Science, University of Hertfordshire, UK

[2]Hertfordshire Partnership University NHS Foundation Trust, Welwyn Garden City, Hertfordshire, UK

[3]Dept. of Psychiatry, School of Clinical Medicine, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK

Keywords: robot model, autonomous robot, embodied artificial intelligence, cybernetics, OCD (obsessive-compulsive disorder), OC-spectrum, model of mental disorder, model design process, model evaluation

ABSTRACT

Computational psychiatry is increasingly establishing itself as valuable discipline for understanding human mental disorders. However, robot models and their potential for investigating embodied and contextual aspects of mental health have been, to date, largely unexplored. In this paper, we present an initial robot model of obsessive-compulsive (OC) spectrum disorders based on an embodied motivation-based control architecture for decision making in autonomous robots. The OC family of conditions is chiefly characterized by obsessions (recurrent, invasive thoughts) and/or compulsions (an urge to carry out certain repetitive or ritualized behaviors). The design of our robot model follows and illustrates a general design framework that we have proposed to ground research in robot models of mental disorders, and to link it with existing methodologies in psychiatry, and notably in the design of animal models. To test and validate our model, we present and discuss initial experiments, results and quantitative and qualitative analysis regarding the compulsive and obsessive elements of OC-spectrum disorders. While this initial stage of development only models basic elements of such disorders, our results already shed light on aspects of the underlying theoretical model that are not obvious simply from consideration of the model.

INTRODUCTION

The growing field of computational psychiatry (Adams, Huys, & Roiser, 2016; Corlett & Fletcher, 2014; Huys, Maia, & Frank, 2016; Montague, Dolan, Friston, & Dayan, 2012; Stephan & Mathys, 2014; Wang & Krystal, 2014) includes within its aims the application of computational techniques to understand, and better treat, human mental disorders. This includes the use of simulations to explore and compare theorized mechanisms for diseases. However, to date, simulations of mental disorders have largely been "disembodied", i.e. the simulation has been divorced from sensorimotor interaction with the environment (see (Yamashita & Tani, 2012) for a rare counter-example). In order to address this gap, we advocate the use of autonomous robots in modeling mental disorders (Lewis & Cañamero, 2017) to augment existing computational psychiatry techniques.

Robot models complement existing biological and computational models in a number of ways. Compared to purely computational models, robots, like animals, allow us to

an open access journal

Citation: xxxxxx

DOI:
http://dx.doi.org/xxxxxx

Supporting Information:
http://xxxxxxx

Received: xxxxxx
Accepted: xxxxxx
Published: xxxxxx

Competing Interests: The authors have declared that no competing interests exist.

Corresponding Author:
Matthew Lewis
M.Lewis4@herts.ac.uk

model complete systems, including a closed-loop interaction with the real environment. Compared to animal models, commonly used in psychiatric research, robot models allow precise operationalization of theoretical models through implementation, increased replicability and control of experiments, and the ability to make controlled manipulations that may not be possible in animals for ethical, methodological, or practical reasons. Such controlled manipulations include, for example, changing the values of specific parameters in the robot model, e.g. in relation to OCD, testing different values of a threshold controlling the tolerance to perceptual errors. Another example would be introducing errors analogous to brain lesions, but in a highly precise and reproducible manner. A third example would be to introduce communication errors between components of the controller, e.g. the addition of noise to signals in a neural network corresponding to errors in top-down/bottom-up communication used by Yamashita and Tani (2012) in a robot model of schizophrenia. In this and other cases, the analogous controlled manipulations in animals may be inaccessible because we either do not know which specific elements or connections to manipulate, or we do not have a technique for manipulating them consistently, or we cannot manipulate them without causing side-effects in other parts of the system (for example, if a drug used also binds in another part of the body). However, when carrying out such manipulations in robots, we should have a clear hypothesis regarding the existence of analogous dynamics or systems linked to the condition that we are seeking to understand in human patients (construct validity – see section Evaluation of the Robot Model (Stage 7)).

Since the use of such robot models is a new area of research, we seek to establish a design framework (methodology, guidelines and evaluation criteria) to guide research (Lewis & Cañamero, 2017). Our motivation in choosing these guidelines is to ensure that we learn from the extensive experience of researchers using animal models, to ground the research in theoretical models, and to guide research towards applications. In this paper, we present the initial development and first experiments for a robot model of obsessive-compulsive disorders following this framework.

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 of disorders, which also includes conditions such as trichotillomania (TTM, pathological hair pulling), pathological skin picking (PSP), body dysmorphic disorder (BDD), and tic disorders such as Tourette's syndrome (American Psychiatric Association, 2013). A cardinal feature of these disorders is the performance of compulsions, which can be defined as repetitive stereotyped behaviors, performed according to rigid rules and designed to reduce or avoid unpleasant consequences but which, as a consequence of the repetition, become a source of distress and functional disability (Fineberg et al., 2018). The behaviorally similar 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. Whether OCPD should be considered within the OC spectrum is an open question (Fineberg, Reghunandanan, Kolli, & Atmaca, 2014).

A number of theoretical models and underlying mechanisms have been proposed for OCD, including cognitive-behavioral models (Shafran, 2005), a cybernetic model (Pitman, 1987), the signal attenuation model (Joel, 2006), exaggerated sense of danger (Apergis-

Schoute et al., 2017), exaggerated sense of responsibility (Mantz & Abbott, 2017; Salkovskis et al., 2000), and bias toward habitual versus instrumental acts (Gillan et al., 2011; Gillan & Robbins, 2014). Of these, our robot model takes closer inspiration from the cybernetic model of Pitman, and the signal attenuation model.

We present and test experimentally a cybernetics- and ethology-inspired autonomous robot control architecture for decision making that can display both adaptive (functional) behavior as well as non-functional decision making that presents similarities with compulsions and obsessions in OCD.

In the remainder of the paper, we first review different types of models of mental disorder, then give an overview of our design process, before illustrating it with our development of a robot model for OCD. We then describe our initial experiments and discuss our experimental results.

MODELS OF MENTAL DISORDERS

Types of Models of Mental Disorders

In previous work (Lewis & Cañamero, 2017) we described four types of model commonly used in research into mental disorders, which we recap these here:

1. A *conceptual model of a mental disorder* is a theoretical construct that links underlying causes (etiology), either proposed or observed, with observed symptoms and correlates. A conceptual model serves as a framework for understanding, and should have explanatory and predictive power with respect to the condition being modeled. There is not necessarily one "true" model, since different models may be complementary, having different scope, emphasis, level of abstraction, or uses.
   A conceptual model may be associated with one or more specific implementations (in the sense of the three types listed below). However, this is not always the case; for example, Pitman's cybernetic model of OCD (Pitman, 1987) had been so far, to our knowledge, a purely conceptual model. A conceptual model without an implementation can nevertheless have applications to guide research into potential treatments (notably, the initial conception of Exposure and Response Prevention as a treatment for OCD was based on a theoretical formulation (Meyer, 1966)), to provide an explanatory framework for observations, or as a theoretical basis for future research – for example, the Cognitive–Energetic Model of ADHD (Sergeant, 2000).
2. An *animal model of a mental disorder* is a non-human animal used to study brain–behavior relations with the goal of gaining insight into, and to enable predictions about, these relations in humans (van der Staay, 2006). Animal models may be induced by genetic manipulation, drugs, or by environmental manipulation. Alternatively, they may be naturally occurring. They have the advantage that they model a complete system (organism and environment) and use a real animal. However, there are limits to how closely a non-human animal can be used to model human mental disorders (Geyer & Marcou, 2002).
3. A *computational model of a mental disorder* is a realization, or partial realization, of a theoretical model in a computer. The field of computational psychiatry includes within its scope the development of computational models of psychiatric disorders (Huys et al., 2016). These models have the advantage that they are highly specified and so any results should be replicable and can be analyzed in detail. However, due to the complexity of implementing such a model, they are typically only partial imple-

mentations (e.g. of a neurological subsystem, as in the model of OCD in Maia and Cano-Colino (2015)) or they work at a relatively high level of abstraction (such as reinforcement and Bayesian learning models – for an overview, see Huys et al. (2016)). In addition, they do not necessarily include any behavioral element, a true closed-loop interaction with the environment, or the effects of contextual and environmental elements.

4. A *robot model of a mental disorder* embeds a computational and hardware realization of a conceptual model in an embodied, interacting robot and its environment. Like an animal model, it models a complete system (agent and environment), but using an artificial agent rather than an animal. While conceptual, animal, and computer models are widely used in research, there has thus far been relatively little use of robot models (one of the few examples is by Yamashita and Tani (2012)). However, robot models share the advantages of computational models in terms of specificity and controllability, while, like animal models, taking into account the agent–environment interaction. We thus advocate the development of robot models of mental disorders, to complement existing models by offering more controllable agents in a complete system, in which theoretical models can be more precisely implemented (Lewis & Cañamero, 2017).

In reality, the different categories of model will not be clear cut. For example the signal attenuation model for OCD (Joel, 2006), outlined below, combines both a theoretical and an animal model.

Design Framework for Robot Models

We have followed the iterative design process shown in Figure 1 for the development of our robot model, with modifications to adapt it for use with robots, which we will describe as we describe the process. This process is based on a design process for animal models of behavioral disorders (van der Staay, 2006; van der Staay et al., 2009), which provides us with a well-established evaluation framework used by the clinical research community, and which is generally relevant to the development and evaluation of embodied models. The design process that we have followed covers all the stages in the design process, starting with a theoretical model, through experimental evaluation and refinement.

We have here refined the process that we proposed in (Lewis & Cañamero, 2017) by changing the "Accept the model" end point into part of the flow, in order to allow the explicit inclusion of "us[ing] the accepted model for clinically focused studies", while also reflecting the fact that a robot model (and indeed, any computational model) is always open to further development and refinement.

In this paper, we will illustrate this process with the development of a robot model for OC-spectrum disorders.

ROBOT MODEL DESIGN PROCESS: PRE-DEVELOPMENT STAGES

Theoretical Model Selection (Stage 1)

The first stage in the development process is to select a conceptual model (or multiple complementary conceptual models) to serve as the basis of the robot model. While a new conceptual model could be created at this stage, there would not then be the opportunity for the wider mental health community to review it before implementation. In practice, since

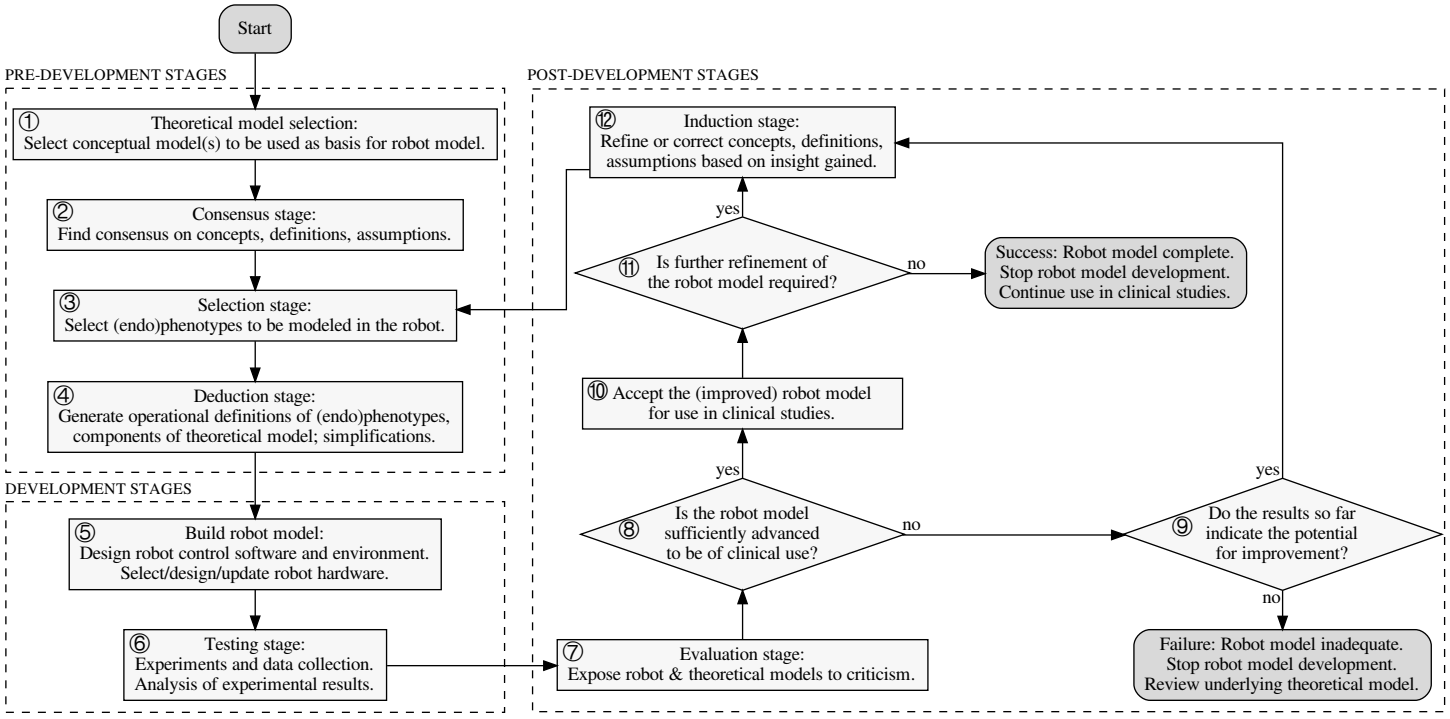

Figure 1. Flowchart for an iterative process for designing a robot model. This is a modified version of the chart from (Lewis & Cañamero, 2017), which is based on, and closely follows, the process described in (van der Staay, 2006; van der Staay et al., 2009). Numbers in circles are to facilitate references to individual steps in the text. Note that, even after the robot model is accepted for clinical use (Stage 10), it is envisioned that robot model development might continue iteratively, and incremental improvements will be made with each loop through the process.

the process of developing the robot model requires the computational and hardware implementation of the model, new components of the model will be created during development.

A variety of theoretical models exist for OC-spectrum disorders (see e.g. (Fineberg, Chamberlain, Hollander, Boulougouris, & Robbins, 2011) for a discussion of various models). Our choice to conceptualize OCD as a disorder of decision making Sachdev and Malhi (2005) and of a specific conceptual model (based on cybernetics, as we explain below) was linked to our (LC and ML's) existing research in robotics, which has extensively investigated decision making (action selection) in autonomous robots from a perspective that is close to cybernetics. Our previous work in decision making (behavior selection) in autonomous robots, inspired by ethology and cybernetics, investigates the adaptive value of decision-making strategies, measured in terms of contribution to maintenance of homeostasis, in motivated and goal-oriented behavior in robots (L. Cañamero & Avila-García, 2007; L. D. Cañamero, 1997; Lewis & Cañamero, 2016). In that work, the overall behavior of the robot was changed in different ways through controlled manipulation of the perceptual element of the perception-action loop, namely through modulation of perceptual properties of incentive stimuli. Alongside adaptive benefits resulting from such benefits, we observed maladaptive behaviors – in particular, excessive persistence in behavior execution – that bore a similarity with decision making problems in OCD and other conditions such as addictions. Given the similarities between our model of decision making, and the cybernetic

(Pitman, 1987) and signal attenuation models (Joel, 2006) of OCD, we selected these as the conceptual models to be used as the basis for our robot model of OCD. As we shall discuss later in the paper, in addition to the behavioral "compulsion" aspect stressed by animal models, both the cybernetic model of OCD and our specific models of motivation also allow us to consider the internal "obsession" aspect of OCD, since elements of the model can be viewed as "thoughts" even when they do not result in action.

Pitman's cybernetic model (Pitman, 1987) takes the cybernetic view of behavior as an attempt to control perception (Powers, 1973). In the cybernetic framework, behavior (of natural or artificial systems) is the result of attempting to correct "perceptual errors". Such errors indicate a mismatch between an "actual" perceptual signal (such as sensed room temperature in the archetypal example of a thermostat) and an internal reference, ideal or "target" value that the system aims to reach (the set temperature that the thermostat tries to maintain). The actual signal can be external, such as in the example of thermostat, although it can also be an internal signal, such as perceived hunger. Pitman specifies that, in the field of control systems theory, the reference signal is an internal signal (e.g. satiety signal after satisfaction of hunger). This does not mean that the target value is fixed, since it may adapt to some extent to adjust to internal or external environmental factors. The core element of such cybernetic control systems is an internal *comparator mechanism* that computes the mismatch between the actual (measured) value and the target value. This difference is conceptualized as an error that provides a signal (called the "error signal") for the system to trigger behavioral output that aims to correct that error (e.g. in the example of the thermostat, activating the heating mechanism). Following this model, Pitman conceptualizes OCD in terms of behavioral control of perception, and proposes that "the core problem in OCD is the persistence of high error signals, or mismatch, that cannot be reduced to zero through behavioral output" (Pitman, 1987, p. 336), for example, an erroneous ever-present perception that the hands are contaminated, leading to compulsive washing that fails to make the erroneous perception disappear. Pitman argues that his model can explain features of OCD such as perfectionism, indecision, need for control, over-specification, and obsessive thoughts, with the presence of the error signal itself being subjectively experienced as a sense of incompleteness and doubt. He further considers three possible sources for the persistent error signal: conflict between multiple control systems, comparator defect, and attentional disturbance.

The signal attenuation model for OCD is a theory-based animal model built on the proposition that "compulsive behaviors result from a deficit in the feedback associated with the performance of normal goal-directed responses" (Joel, 2006). In the associated experimental animal model, compulsive behavior in produced by the attenuation of the informational value of an external signal (e.g. light or sound) that has been linked, by training, to the successful execution of some action (e.g. lever pressing for food). A more generalized view of this model would also include internal signals, such as interoceptive signals for satiety after eating or drinking. Indeed, internal signals are important in goal-directed and motivated behavior (Damasio, 2010; Frijda, 1986; Lehner, 1996; Panksepp, 1998; Pessoa, 2013), e.g. to provide to provide targets, and "stop messages" or to monitor performance. However, internal signals are are not normally accessible (for technical, practical or ethical reasons) in studies involving animal models, which must resort to the use of external signals and their association (typically through learning) with externally observable behavior.

We therefore select the cybernetic model of Pitman, and the signal attenuation model (as a theoretical model, rather than its specific implementation in animals) as the basis for

our robot model. These are brought together within the framework of our motivation-based robot controller. We will implement our robot model of OCD using an internal signal deficit (faulty interoception); this is something that is much simpler to do in robots than in animals, due to our more complete control of the robot's internal decision-making and sensing mechanisms. The internal signal deficit falls within the category of "comparator defect" in Pitman's possible sources for the error signal.

Consensus Stage (Stage 2)

This stage seeks conceptual clarity regarding the selected conceptual model, and precision in the use of its associated notions and principles. We refine and seek consensus, typically on concepts, criteria, definitions and assumptions underlying and associated with the conceptual model of the previous stage.

From our selected cybernetic and signal-attenuation conceptual models, the key notions with respect to OC-spectrum disorders that are most relevant for our robot model are:

- *Compulsions*. According to the classic text of Fish (Fish, Casey, & Kelly, 2008), compulsions are obsessional motor acts that may result from an obsessional impulse or thought. To attempt to clarify this, we will consider a behavior to be compulsive if it is executed repetitively and persistently, even though it might not have a clear function, or it could even be maladaptive, or unwelcome to the individual. We note that some habits may fall under this definition of compulsive behaviors; however, the main difference between them might reside in the context in which they are executed.
- *Compulsive grooming*. In research using animal models, compulsive self-grooming is widely used to research OC-spectrum conditions. It is induced in mouse models by gene knockout, and its link with human OC-spectrum disorders is supported by the similar responses to pharmaceutical interventions (Camilla d'Angelo et al., 2014). Grooming is considered related to the human conditions of trichotillomania (TTM, pathological hair pulling), pathological skin picking (PSP) due to the high face validity. In this case, we consider grooming behavior to be compulsive if it occurs to the extent that it either directly damages the animal, or that it causes the animal to neglect other needs.
- *Obsessive thoughts*. Obsessive thoughts are a defining feature of OCD. According to Fish, obsessions are thoughts that persists and dominate an individual's thinking, despite the individual's awareness that the thought is without purpose or no longer relevant or useful.[1]
- *"Stop signals"* are internal or external signals that indicate that a goal has been achieved, a need satisfied, or a behavior successfully executed. Several models of OCD, namely the cybernetics and signal attenuation models, postulate problems with "stop signals" linked to compulsive behavior. In the cybernetic model (Pitman, 1987), an error signal is present, such that, when it becomes zero, signals that the behavior that was being executed to correct the error can stop. In the signal attenuation model, a sig-

[1] Note that the "obsessive thoughts" present in OCPD (which is not always considered as belonging on the OC-spectrum) are not obsessive in the sense defined here, since they are viewed by the individual as having a worthwhile purpose. Hence this notion of obsessive thoughts helps to distinguish between two conditions. However, at this stage, our robot model does not have any way of assessing whether the repetitive thoughts are worthwhile or not.

nal indicates the successful execution of a behavior; compulsive behavior then results from an "attenuation" of that signal, which weakens the perception of the behavior's success, and therefore that it can stop. The signal attenuation model is typically presented in the context of an experimental paradigm (Joel, 2006) in which animals are trained on an external signal and this signal is "attenuated" by reducing its value as a signal. However, we will consider it more generally, and in our robot experiments the equivalent of the stop signal will be an internal one.[2]

Selection Stage (Stage 3)

At this stage, we select the (endo)phenotypes of interest for our model. These may be behavioral or internal phenotypes that we generate explicitly, or that we believe may be generated as a consequence of how the model works.[3]

Since we were starting the development of our model, we choose to model one of the the simpler conditions in the OC-spectrum: compulsive self-grooming. While compulsive self-grooming is our behavioral phenotype of interest, we are also interested in endophenotypes, in particular, obsessions – a major subjective symptom of OCD. The endophenotypes that we can study depend on the nature and interaction dynamics of the robot controller. In our case, our robot controller uses competing motivational systems that vary over time as a function of the robot's interaction with the environment and the dynamics of its embodiment. In this model, we can use these motivational states as an indicator of obsessions. Such subjective symptoms are not easy to analyze in animal models, where access to the internal state is limited. The list of phenotypes of interest may expand on subsequent iterations.

Deduction Stage (Stage 4)

At this stage, we create operational definitions of concepts to be used in this iteration of the development. In some cases, simplifications of concepts may be required.

First we need to describe briefly how our robot model will work. We will have an autonomous mobile robot that tries to survive in an environment, making decisions about how to use the resources available to satisfy its needs. It will be endowed with some internal physiological variables (e.g. energy), the values of which will change over time as the robot interacts with the environment, and which may fall out of the range of permissible or viable values (Ashby, 1960), resulting in the robot's "death". The robot will also be able to self-groom through interaction with appropriate objects in the environment. We will analyze the robot's behavior and performance in terms of metrics to measure "viability" and "wellbeing" of the robot, as well as statistics about its behaviors. These metrics (to be described in sections Metrics and Testing Stage: Analysis of Experimental Results (Stage 6b)) will be calculated from the above-mentioned internal physiological variables that constitute the internal state of the robot.

[2] Readers familiar with the Stop Signal Paradigm (Verbruggen & Logan, 2008), should note that this paradigm requires that the stop signal takes time to be processed; however, in this paper our analog to the stop signal will propagate instantaneously, but can have different strengths.

[3] Note that, compared to our earlier presentation of this design process (Lewis & Cañamero, 2017), we have moved the Selection stage to after the Consensus stage, so that the conceptualization of the (endo)phenotypes is consistent and clear before selecting the focus of our robot model.

In this context, taking the concepts from the Consensus Stage, we refine them (for this iteration) as follows:

- Adaptive (maladaptive) behavior in the robot is behavior that positively (negatively) affects the performance of the robot, as measured by the viability and wellbeing metrics.
- Compulsive grooming is repeated self-grooming to the extent that it is maladaptive.
- Obsessions are persistent states in the robot's internal decision-making process that have no benefit, either because they cannot be acted upon, or because acting on them will have no benefit in terms of the robot's viability or wellbeing.
- The error signal of a physiological variable is the mismatch (difference) between the current value and the target (ideal, reference) value.
- The perceived error signal of a physiological variable is the robot's "sensed" difference between the current value and the target value (in our case, the perceived error signal may be different from the actual error signal (section Modeling Compulsive Behavior), and our robot's action selection code uses the perceived value).
- Signal attenuation is a decrease in the strength or salience of an internal or external cue. In the signal attenuation animal model, this cue is an external cue to indicate that a behavior has been successfully executed, and the next stage in a chain of behaviors can be started. However, in our case we use an internal variable that can have different target values under different conditions, leading to different error signals, and hence different signals that a behavior (grooming) has had sufficient effect.

ROBOT MODEL DESIGN PROCESS: DEVELOPMENT STAGES

Build robot model of OC-Spectrum Disorders (Stage 5)

To begin this section, let us highlight the features and advantages provided by an (embodied) robot model versus a computer simulation of an agent. In an embodied robot model, the external environment (and the way it is perceived by the robot) is as important as the internal controller in producing the robot's behavior. The environment, in addition to posing specific decision making problems to the robot, provides the context through which the robot's behavior links back to and modifies its perceptions, closing the perception-action loop (Brooks, 1991a; Pfeifer & Scheier, 2001; Powers, 1973). Given the same behavior control software and the same internal state of the robot, the behavior of the robot might be completely different depending on factors concerning its relation with the environment, such as: what is happening in the environment at a particular time, its ambiguity and unpredictability, what the robot perceives of it, imprecisions in the perception-action loop (e.g. in the case of robots, the potential "noise" coming from sensors or actuators), how the robot can act on the environment and how the robot and the environment interact, how what is happening in the environment affects the actions of the robot, the opportunities for interaction that the environment "affords" to the robot, the place of the robot in the ecological niche, etc. The dynamics of such highly complex interactions cannot be fully modeled in a simulator, since the complexity of the real world and its effects on an agent cannot be fully modeled. Since we are interested here in the dynamics of the pathology, and this is something that occurs in interaction with the real world, we advocate the use of a robot model situated and in interaction with the physical world.

Robot hardware For our initial model, we selected a simple robot, well suited to prototyping and research of an exploratory nature: the Elisa-3[4] (Figure 3). This is a small, round two-wheeled Arduino-based robot, 5cm in diameter and 3cm in height. It is equipped with a ring of eight infrared (IR) "distance" sensors with a range of approximately 5cm, and four downwards-pointing IR "ground" sensors. These sensors provide the robot with a rudimentary (coarse and noisy) capability to detect both the proximity of objects around it, and dark and light areas on the ground. It additionally has radio communications to receive and transmit messages with a PC, which we use to log data for quantitative analysis of results. Finally, it has colored LED lights on the top of the robot, which we used to visually signal its internal activity.

Since this robot has limited capabilities for manipulation and perception of external objects[5], we model grooming by having it "rub" its side sensors against objects in the environment to improve its (simulated) "integument": the state of its external surface, analogous to the state of an animal's fur or feathers.

Environment For the purposes of data collection and analysis, we have placed the robot in a small walled area containing objects appropriate to the sensorimotor capabilities of the robot. Our environment had to support both "healthy" and pathological behaviors. We therefore placed in the environment a number of resources – "energy sources" (light patches on the floor) and "grooming posts" (plastic pipes) – that could provide the means for the robot to satisfy its survival-related needs (energy) and its other needs (grooming for maintenance of integument), but which could also provide the opportunity for pathological behavior.

An internal "integrity" variable, keeps track of the (simulated) physical integrity of the robot. It decreases following collisions and other types of contact with objects (detected by the distance sensors): these include the arena walls and the grooming posts placed within the arena. If this damage causes the robot's integrity to fall to zero or below, the robot will "die" and stop in place. In the absence of collisions, the integrity would slowly increase as the robot "heals". In the architecture of the robot, the integrity variable is linked with the robot's motivation to avoid objects, including the grooming posts, providing an internal conflict with the motivation to groom. This gives both a cost to the grooming behavior and a "cue" to stop. This element of the architecture allows us to investigate the extent to which the grooming behavior is compulsive, specifically, the extent to which it continues even though it directly damages the robot (see the Consensus Stage (Stage 2) section, compulsive grooming).[6]

In healthy (adaptive) decision making, the robot will alternate between grooming (or seeking a grooming resource) and feeding (or seeking an energy resource). In the compulsive behavior situation, the robot will groom to the extent that it adversely affects its

[4] http://www.gctronic.com/doc/index.php/Elisa-3

[5] Future iterations of our model may use different robot platforms, as we use it to execute more complex behaviors. However, starting with a simple platform means that initial development does not require a complex controller. With the rise of 3D-printing it will be possible to use robot platforms that are highly customized to the application.

[6] In the future, the inclusion of physical damage from grooming may allow us to link to research on OC-spectrum disorders that has linked them with an increased tolerance of pain (Hezel, Riemann, & McNally, 2012).

Table 1. The robot's physiological variables.

Variable	Fatal limit	Ideal value	Maintenance
Energy	0	1000	decreases over time; increases when the robot consumes from an energy resource
Integrity	0	1000	decreases on contact with objects; increases over time as the robot "heals"
Integument L	none	1000	decreases over time; increases when the robots left side passes close to a grooming post
Integument R	none	1000	decreases over time; increases when the robots right side passes close to a grooming post

survival, either because its energy level falls too low, or because it damages itself though contact with the grooming post.

Robot Model of Obsessive-Compulsive Disorders

As we have seen, the theoretical cybernetic model that we have selected proposes that "the core problem in OCD is the persistence of high error signals [...] that cannot be reduced to zero through behavioral output". To operationalize this, our architecture will act on the basis of the perception of internal error signals, combined with (perceived) external cues. By manipulating internal parameters within the controller to create cases where the error signal remains present, we can then test and explore the theoretical model. To facilitate systematic analysis of experimental results, we test the robot in a "two-resource problem" (Spier & McFarland, 1997), used in ethology and robotics as the simplest decision making, or action selection, scenario. As its name suggests, in this scenario, an agent (animal or robot) must autonomously decide which of the two resources available in the environment it should consume in a timely fashion in order to satisfy its survival-related needs successfully. In order to focus on the dynamics of the perception-action loops that are proposed as the "core problem" in OCD, our robot does not include other elements, such as memory, learning or map building.

Software Behavior Control Architecture The specific robot model and action selection architecture that we have implemented draws on our previous work on motivation-based robot behavior controllers (L. Cañamero & Avila-García, 2007; L. D. Cañamero, 1997; Lewis & Cañamero, 2014, 2016), while also being inspired by animal models of OC-spectrum disorders.

In the robot architecture used in this study, four competing motivations guide the behavior of the robot. These motivations are urges to action determined by a combination of the four corresponding internal homeostatically-controlled "physiological" variables, that provide the robot with "needs", and by the robot's perception of the environmental resources that can be used to manage those variables. The decision making behavior control software provides the robot with strategies to prioritize and satisfy these needs.

An overview of the behavior control (also known in the literature as action selection) mechanism is shown in Figure 2. We describe the components of the architecture (the high-level design of the software) in the following subsections.

Physiological Variables Our robot has four homeostatically-controlled physiological variables shown in Table 1: energy, integrity, and two integument variables (one for each side). The physiological variables take values in the range [0,1000], with 1000 being the ideal

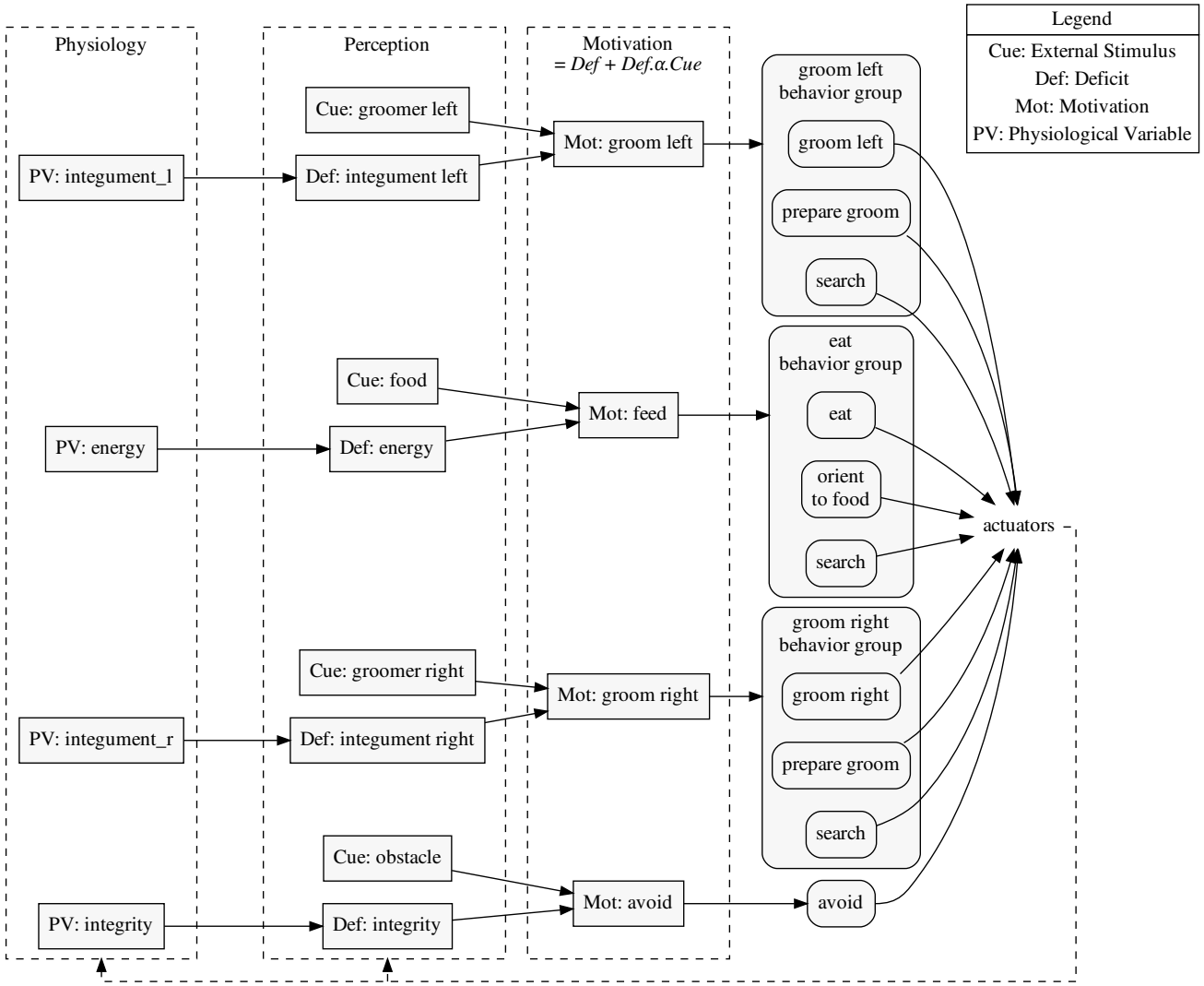

Figure 2. An overview of the action selection mechanism for our robot. Rounded boxes represent individual (potentially nested) behaviors, while square-cornered boxes represent other internal components. The actions of the actuators result in changes in the environment and the robot's physiology, which is fed back to the robot controller via the robot's perceptions. Motivations are updated and new behaviors selected every action selection loop (10Hz).

value in all cases. The variables change both over time and as a function of the robot's interactions with its environment, reflecting the current state of the robot. Following a model of homeostatic control, the difference between the actual value and the ideal value of each variable generates an error signal indicating the magnitude of the mismatch (in this case, deficit).

Two of the physiological variables (energy and integrity) have a fatal limit of zero (the robot dies if the value falls to zero). The two other variables, related to "integument", can be thought of as analogous to an animal's fur or feather condition: something that needs to be maintained for viability (e.g. waterproofing, efficient flight), but which doesn't directly cause death if it falls too low. In our robot implementation, low values of the

integument variables have no physical consequences on the robot (e.g. it doesn't affect its physical integrity or its travel speed or any other aspects), but they will trigger a motivation to maintain the variable within a good range of values (correct the error between the actual and ideal values of this variable) by grooming.

Sensors, Cues and Motivations The robot uses its infrared distance sensors and ground sensors to detect obstacles, grooming posts and energy resources in the environment. These correspond to environmental cues or incentive stimuli that influence the motivational states of the robot to avoid (obstacles), groom (at grooming posts) or consume (energy resources). The numerical size of the perceived cue is in the range [0,100] and is determined by the sensed distance of the obstacle, or by the color detected by the ground sensors for the energy resources and grooming posts.

Following a classical model in ethology, we use the long-standing concept of motivational states (Colgan, 1989), defined in terms of the drives set by the deficits or errors of the internal variables, combined with external environmental cues (incentive stimuli). Our robot has four different motivations, each linked to the satisfaction of a physiological variable (see Figure 2). The internal drives and the external incentive cues combined, provide a level of intensity to each motivation at each point in time, which reflects its relevance to the current situation. The motivational intensity is calculated according to the formula proposed in (Avila-García & Cañamero, 2004) (modified from a classical formula in ethology (Tyrrell, 1993, p. 139)):

$$motivation_i = deficit_i + deficit_i \times \alpha \times cue_i \quad (1)$$

Where $deficit_i$ (the error signal) is the difference between a variable's current value and its ideal value as perceived by the robot (see section Modeling Compulsive Behavior), cue_i is the size of the corresponding cue, and α is a scaling factor to scale the size of the exteroceptive component. In our experiments, α will equal 0.05. This value was empirically determined in pre-trials to allow the robot with a realistic target value (non-pathological or baseline condition 1 in our experiments below) to be able to have enough persistence in satisfying its needs to be able to survive in the environment.

As the values indicating the intensity of the motivations change over time, depending on the external and internal perceptions, at certain points the motivation with the greatest intensity (i.e. the most pressing need) will be overtaken. This will result in a change of motivational state, and hence of the behavior executed to satisfy it, and it could be viewed as an analog for a "stop signal" for the current behavior.

Behaviors The robot has a number of discrete behaviors of different types (rounded boxes in Figure 2). The execution of some of these behaviors allows the robot to directly satisfy its motivations, and hence correct the errors of the physiological variables, whilst other behaviors allow the robot to "search" (move about the environment avoiding objects) for the resources required to satisfy these needs. Let us note that the robot will only move around the environment or execute a behavior if at least one of its motivations is active.

We have grouped the behaviors into four "higher level" behavioral subsystems, each one linked to a motivation: groom left group, groom right group, eat group, and avoid. These behavioral subsystems (except for avoid) are composed of smaller simpler behaviors, which can be executed independently, simultaneously or sequentially depending on

the state of the robot and the external stimuli detected. For example, the eat behavioral group is composed of the consummatory (goal-achieving) behavior "eat", which is executed if the robot is hungry and food is detected and located at the mouth, "orient to food", which is executed if the robot is hungry and food is detected nearby, and "search (for food)" is an appetitive (goal-seeking) behavior that will make the robot wander around the environment until food is detected. Searching behaviors (for an energy resource or a grooming post) involve the robot wandering around the environment (i.e. travelling in random directions, while avoiding objects), typically until the incentive stimulus of the motivation that triggered the search is found. Searching does not involve any knowledge of the environment on the part of the robot, and does not occur in any particular direction. The only information about energy resources and grooming posts that the robot uses in this search, is the ability to recognize them when it encounters them.

Action Selection The robot controls its behavior as follows: At each time step (100ms), the robot recalculates its four motivations and sorts them from highest to lowest. This order determines the order in which it prioritizes the satisfaction of its physiological variables in that action selection step. To satisfy the motivations, we use a slightly modified "Winner-Take-All" action selection policy, as follows. The robot will try to satisfy the highest ranked motivation (the "winner motivation") first. To do so, the winner motivation triggers the behavioral subsystem linked to it, and one or more of the simpler behaviors that constitute this subsystem (nested rounded boxes in Figure 2) are executed, depending on whether the preconditions for their execution (e.g. in the case of the eat behavior, presence of food detected) are met. If, while satisfying the winner motivation, a behavior that satisfies a lower ranked motivation can also be executed, then it will be executed. This means that two behaviors can sometimes be executed simultaneously. Our robot can execute two behaviors simultaneously only if the two behaviors use distinct sets of actuators. For example, the "orient to food" behavior, which allows the robot to approach and stop at an energy resource, uses the wheels, while the behavior to consume the resource uses the virtual "mouth", and thus the two behaviors can be executed concurrently. This allows the robot to consume an energy resource as it aligns itself with the resource. In certain cases, two behaviors can execute due to different motivations, for example, the eat behavior can be executed opportunistically as the robot passes over a resource while searching for a grooming post, as the mouth actuator is not otherwise occupied.

Note that, since the grooming posts are obstacles that the robot can collide with, they will also act as a cue for the avoid behavior. Whether the avoid behavior is actually executed depends on the intensity of the motivation to avoid, which depends on the values of the cue and the robot's integrity.

Modeling Damage and Grooming Damage to the robot (e.g. through collisions) and the effect of grooming on the robot's integument are implemented using the IR distance sensors. Damage and grooming use independent mechanisms (summarized in Figure 3) designed with the goal that interaction with environmental objects can be potentially beneficial or damaging to the robot: grooming involves a small possibility of damage, but not so much that a normal amount of grooming risks serious damage to the robot. The various constants in these calculations were determined empirically to meet these design goals.

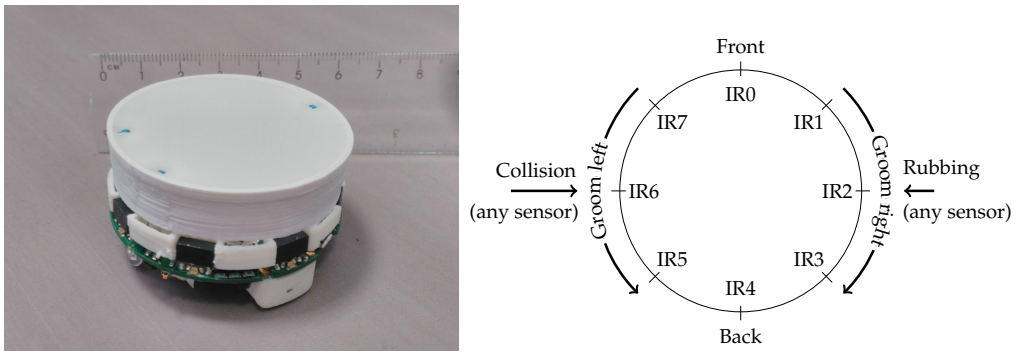

Figure 3. Left: An Elisa-3 robot, viewed from the front/left. Right: A diagram of the Elisa-3's infrared distance sensors (top view). Arrows indicate how the sensors are used to detect grooming and damage from collisions and sustained rubbing.

To calculate damage, the distance sensors are checked every 50ms, and compared to the previous values. The calculated damage is subtracted from the current integrity. Two types of damage are possible: collisions, and sustained rubbing:

- Collisions: if the closest IR reading crosses a "touch" threshold (a value of 850, corresponding to an object approximately 3mm from the robot), then a "collision" is deemed to have occurred, and a value for the damage is calculated depending on the size of the change in the sensor readings that have crossed the threshold. A maximum value of 100 is applied to this type of damage, to stop a single unlucky collision killing the robot in one blow.
- Sustained rubbing: when the IR sensor values maintained a value over the threshold of 900 (indicating a very close object) then a constant value of 2 damage is applied (hence a maximum of 40 per second).

To implement grooming, the distance sensors are checked every 100ms, and compared to previous values. Two sets of sensor are used: (IR1, IR2, IR3) for the right side, and (IR5, IR6, IR7) for the left side, corresponding to the two integuments. If a sensor value indicates a close object (a value above 150) and the value has increased since the previous reading, while the value of the adjacent sensor towards the front of the robot has decreased (indicating the movement of a grooming post from front to back of the robot) then a "stroke" counter is incremented, indicating movement in the "correct" direction (front-to-back). Conversely, a sensor indication of a movement in the "wrong" direction (back-to-front) is considered as a "anti-stroke" and decrements the stroke counter. The overall value of the counter indicates whether the overall movement on one side is a stroke (> 0), or an anti-stroke (< 0). If a stroke has occurred, then the actual value of the corresponding integument (left or right) is increased by 20 times the count value; if an anti-stroke has occurred, then the actual value of the corresponding integument is decreased by -5 times the (negative) count value.

Modeling Compulsive Behavior Following the signal attenuation model, in this paper we model compulsive behavior by manipulating the robots' perception of the internal errors linked to its physiological variables. More concretely, we manipulate the robot's perceived

ideal ("target") value for the integuments. This will affect the decision making (action selection) process in the calculation of the "error" in Figure 2 or the deficit in equation 1.

To model "typical" or "healthy" decision making, we consider the case where the perceived ideal value is equal to the real "perfect" value (i.e. 1000, which is as far from the fatal limit as possible). This value can be achieved, and so is a "realistic" target. However, as we have seen in previous work, a robot will typically stop attending to a need before the related physiological variable reaches its ideal value, due to competition from other needs. The point when the value of the active motivation is overtaken by another motivation (which consequently becomes the active motivation) can be thought of as the robot receiving the "stop signal" for that behavior (although in our model it might be more accurately thought of as an "attend to another need and switch behavior" signal).

To model "pathological" conditions, we consider values for the perceived ideal value that are not achievable, even theoretically: in this model, values greater than 1000, which is the maximum possible value the variable can take. Since these values are not valid values for the integument, having such a target value can be thought of as a "perceptual error" or distorted perception in the robot.

As outlined in the section on Deduction Stage (Stage 4), if the signal attenuation model of OCD holds, we would hypothesize that this manipulation, analogous to attenuating the signal for successful grooming as we increase the target value, would result in an exaggerated (more intense) motivation to perform the selected behavior. In turn, this increased motivation would out-compete the other needs, and thus produce an increase in the performance and perseverance of the grooming behavior. If this effect is sufficiently large, we would expect it to be maladaptive, and this would be measurable by at least some of our metrics.

Testing Stage: Experiments and Data Collection (Stage 6a)

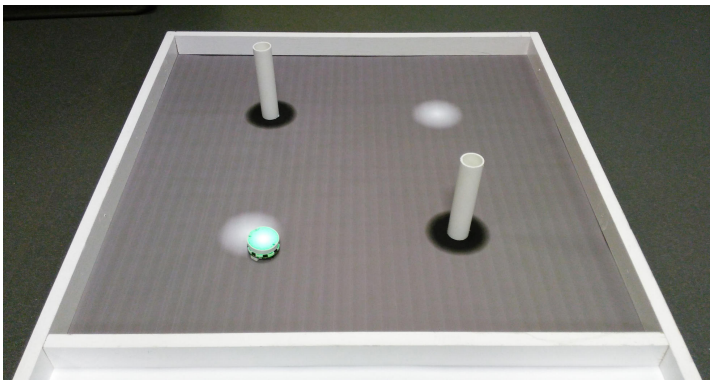

Figure 4. The 80cm × 80cm environment used in the experiment. Here, the robot is feeding at an energy resource (white patch) while the grooming posts (white pipes) stand on the black patches.

Methods The experimental setup is shown in Figure 4. It consists of a 80cm × 80cm square surrounded by 45mm high wooden walls that can be detected with the robot's distance sensors. The floor is covered in paper, which is printed gray, except for two light areas and two dark areas, which can be detected by the robot's ground sensors, and that

indicates the presence of food and grooming resources, respectively. Two 35mm diameter white plastic pipes are fixed in place in the centers of the dark areas to be used as grooming posts.

We conducted twenty runs in each of the following conditions:

1. Realistic target values: Perceived ideal values for integuments = 1000
2. Mildly unrealistic target values: Perceived ideal values for integuments = 1100
3. Highly unrealistic target values: Perceived ideal values for integuments = 1200

Note that conditions 2 and 3 correspond to target values (perceived ideal values) that are not achievable, even in theory, because they lie outside the range of the variables (see section Modeling Compulsive Behavior). In these conditions, an error is always perceived, even in the absence of a "real error" (a mismatch between the actual value and a realistic ideal value within the range of the variable, which is 1000). This "distorted perception" of the target value gives rise to "perceptual errors" that cannot be corrected through behavior or by interaction with the environment because the target values lie outside the range of the variable, and therefore the urge to correct it is always present. These conditions fall under the category of "comparator defect" in Pitman's possible sources of a persistent error signal.

The numerical values for conditions 2 and 3 were empirically determined following some informal pre-trial runs. In these, we observed that the value of 1200 resulted in highly persistent grooming: the robot would frequently groom until it died, so it was selected as the most extreme value to test; while the value of 1100, halfway between the baseline condition and our extreme value, showed very different behavior, with the robot stopping grooming before it died.

On each run, the robot's physiological variables were initialized to the middle value of the range (500 out of 1000) for energy and both integuments, so that the robot would need to work to maintain all these variables, which decrease over time if not actively maintained. For the integrity variable, an initial value of 900 (out of 1000) was chosen to allow the robot the approach grooming posts and maintain its integument, rather than starting in a "half damaged" state and being over-motivated to avoid objects. The robot was started at the center of the arena, equidistant from all four resources (see Figure 4) facing directly towards one side of the arena in one of four alternating directions (labeled "north", "south", "west" and "east"). The alternating direction was done in order to reduce any bias that the initial direction might impose, since this may influence which resource a robot would encounter first. Runs lasted for six minutes each, or until the robot died. The values of its physiological variables, motivations, sensor values and the currently executing behaviors were recorded every 250ms and transmitted to a PC via its radio link.

Select data to be collected During our experiments we will need to collect data to evaluate the adaptive or maladaptive value of the robot's decision-making process. In terms of OC-spectrum conditions, to compare the balance between the satisfaction of different needs, we need to record the values of the physiological variables, the values of the motivations, and the behaviors that the robot is currently executing. To allow post-hoc examination of the robot's behavior, we additionally record its sensor readings and wheel speeds.

Metrics We evaluated the performance of the robot in each conditions and run using four types of metric:

Table 2. Experimental results. For each condition, from left to right: metrics for number of deaths, arithmetic and geometric wellbeings, variance of physiological variables (which can be thought of as the robot's "physiological balance"), percentage of the robots' lifetime spent grooming and eating, and the percentage of the robots' lifetime during which either of the two integument values was zero. The mean wellbeings and variance have been calculated by taking the means over the lifetime for each "robot" (run), and then calculating the mean of the twenty values in each condition. The percentages in the last three columns have been calculated by concatenating the lifetimes of the robots in the twenty runs in each condition, and calculating what percentage of this time was spent grooming etc.

Condition	No. of deaths	Mean arit. wellbeing	Mean geom. wellbeing	Mean variance	%-age time grooming	%-age time eating	%-age time with zero integument
1	2/20	560.9	456.5	55438	34.5	21.3	13.8
2	3/20	603.2	501.1	57435	39.4	20.6	12.5
3	19/20	556.5	344.9	101264	64.9	13.0	27.0

1. *Survival related*. Specifically: Death rates (the number of robots that died during the run), and duration of life for each run (up to 6 minutes).
2. Metrics relating to the regulation of physiology (*wellbeing*, *physiological balance*, *maintenance of physiological variables away from dangerously low values*, *maintenance of integument*). These give a measure of the success of a living robot in managing its physiological variables as a result of its decision making, either at a specific time, or over its lifetime. These help to compare the performance of robots which do not die. They were calculated from the recorded values of the physiological variables.
3. *Behavior of the robot*. The behaviors that the robot executed are included in our logged data, so we can use this record to compare different robots' behavior without needing to resort to external observation.
4. *Motivational balance*. Since our robot controller is based on the four motivations defined in the Sensors, Cues and Motivations section, we can use this internal information to analyze the robots' motivational balance, which reflects how much time it spends attending to each of its physiological needs.

We provide the mathematical definitions of these measures and use them to evaluate the results of our experiments in the following section.

Testing Stage: Analysis of Experimental Results (Stage 6b)

We now present the experimental results and analyze them in terms of the above metrics.

Death Rates, Duration of Life Death rates for each condition are shown in Table 2 (first column). Let us first consider condition 1 (realistic target values). The two condition-1 robots to die both survived 49 seconds, and the runs were very similar. Both of them found a grooming post in this time, spent some time grooming, and then left due to motivation to find an energy resource. After this, they both encountered a second grooming post, and although they both groomed, they did this for only about one second before leaving again in search of an energy resource.

In condition 2 (mildly unrealistic target values), three robots died. Since we recorded internal data for the robots, including their motivations, we can examine what happened, including why they made their decisions. Let us look at what happened during the life of the three robots that died:

- The first condition-2 robot to die survived 52 seconds. It had found a grooming resource soon after starting, and spent 15 seconds grooming before leaving. It soon found another grooming post, and remained grooming for approximately 9 seconds. It left the post with only 6 seconds worth of energy and failed to find an energy resource in this short time. We remark that the integument that the robot attended to was low (only 24 greater then the energy) when the robot prioritized it over the similarly low energy, meaning that the robot had two pressing competing needs.
- The second condition-2 robot to die survived 297 seconds. After 200 seconds, both of its integuments had fallen to zero, and when it found a grooming post, it stayed there grooming for approximately one minute, increasing its integuments while its energy level fell. By the time it left the post, it had approximately 15 seconds worth of energy left, and it did not find an energy resource before dying. As in the first case, the low integuments meant that the robot had multiple pressing needs.
- The third condition-2 robot to die survived 124 seconds. After finding its first grooming post and grooming, it left with approximately 30 seconds worth of energy and was wandering the arena to find an energy resource. However, during this time it found another grooming post and opportunistically groomed for 10 seconds. When it left the second grooming post it had less than 10 seconds worth of energy with which to find an energy resource.

Of the nineteen deaths in condition 3 (with the highly unrealistic target values), seventeen occurred due to the energy falling to zero while the robot was grooming. In all of these cases, the integument in question had already reached its ideal value (i.e. its maximum value of 1000), so further grooming was not achieving anything. In the remaining two cases where the robot died, these were also caused by the energy reaching zero. In both cases, the robot had stopped grooming within the last 5 seconds, and was now wandering – in one case motivated to find an energy resource, in the other case motivated again to find a grooming post. The mean survival time for condition 3 was 134 seconds (including the robot that survived).

Wellbeing In order to evaluate the robot's current state, in terms of its physiological variables, we used metrics that we call "wellbeing" (Lewis & Cañamero, 2016), which provide the average level of all four the physiological variables at each point in time. Intuitively, wellbeing gives an indication of the robot's internal "health" at a point in time, with high values indicating good health (the physiological variables are high, close to the ideal value), and low values indicating poor health (the physiological variables are low, close to the fatal limit).

We calculated two different wellbeing metrics by taking the arithmetic and the geometric means of the physiological variables at each sample time, giving, respectively, the arithmetic wellbeing and geometric wellbeing. Both means are unweighted, i.e. in both means, the four components contribute equally to the final value. However, the value of the geometric wellbeing is more strongly affected by those physiological variables that have values close to the critical value of zero – which is also and fatal limit for energy and integrity – and that, for this reason, can be considered the most pressing. We also calculate the more broadly used arithmetic wellbeing because it gives a more familiar and intuitive way of calculating the average, since it gives the "middle" value.

Applying these metrics to our experimental data, the mean values of each wellbeing metric over all the runs in each condition are shown in Table 2 (second and third columns).

In both the arithmetic and geometric cases, these values were calculated in three steps. First, in each run, in each of the conditions, we calculated the wellbeing of the robot at each "time step" (each point in time for which data was collected). Second, we calculated the mean wellbeing over the lifetime of the robot in each run (twenty runs per condition, sixty in total). Finally, we calculated the overall mean for each condition as the mean of the twenty "lifetime means" (one per run) from the previous step. The mean values of the geometric wellbeing for each run are shown in Figure 5.

There was a statistically significant difference for the arithmetic wellbeing between conditions (ANOVA, $p = 0.020$), with Tukey HSD post-hoc analysis showing that condition 3 (highly unrealistic target values) differed from condition 2 (mildly unrealistic target) ($p < 0.03$), but no statistically significant differences between other pairs of conditions. Turning now to the geometric wellbeing, there was again a statistically significant difference between the different categories (ANOVA, $p = 2.5 \times 10^{-4}$). In this case, Tukey HSD post-hoc analysis showed that condition 3 (highly unrealistic target values) differed from both condition 1 ($p < 0.005$) and condition 2 ($p < 0.001$), while conditions 1 and 2 did not differ significantly from each other.

The fact that the difference in the geometric wellbeings was statistically more significant than the difference between the arithmetic wellbeings (both in terms of the p-value for the ANOVA, and there being a significant difference between conditions 1 and 3) illustrates a desirable property of the geometric wellbeing metric: the larger effect of small ("critical") physiological variables on the overall value. This means that the geometric wellbeing is strongly affected by those variables with large errors, reflecting more accurately the significance of poorly maintained variables. Conversely, for the arithmetic wellbeing one well-maintained variable can cancel out the effect of a poorly-maintained variable, so in an extreme case, a robot could maintain one variable well, but die quickly due to neglecting a survival-related variable, and still have a high arithmetic wellbeing over its lifetime.

In summary, as expected, a highly unrealistic target value (condition 3) is disadvantageous in that it results in a lower geometric wellbeing than either the realistic target value (condition 1) or the mildly unrealistic target value (condition 2). However, contrary to what we would expect, a mildly unrealistic target value (condition 2) is not disadvantageous compared to the realistic target value (condition 1): the trend, although not statistically significant, is that our chosen mildly unrealistic target value results in a higher mean wellbeing than a realistic target value, and so may be advantageous, as measured these metrics. This can be viewed as an advantage of a more cautious decision-making strategy for the management of the integument variables: for the save value of integument the motivation to groom is higher. Even if the advantage of condition 2 does not hold true in further experiments, our results indicate a nonlinear response of the wellbeing metrics to the changing perception of the target value.

Physiological Balance/Variance of the Physiological Variables We calculated the "physiological balance" as the variance of the four physiological variables at each "point in time" (i.e. the variance of the four values at each sampling time, rather than for the entire series of values for each individual variable). Intuitively, this gives a measure of whether, at that point in time, the robot has managed the physiological variables evenly – the four variables have similar values, so the balance (their variance) is low – or whether the robot has kept some variables high while allowing others to fall – the four variables have a wide range of values, so the balance (their variance) is high. A high value can be thought of as a "poorly

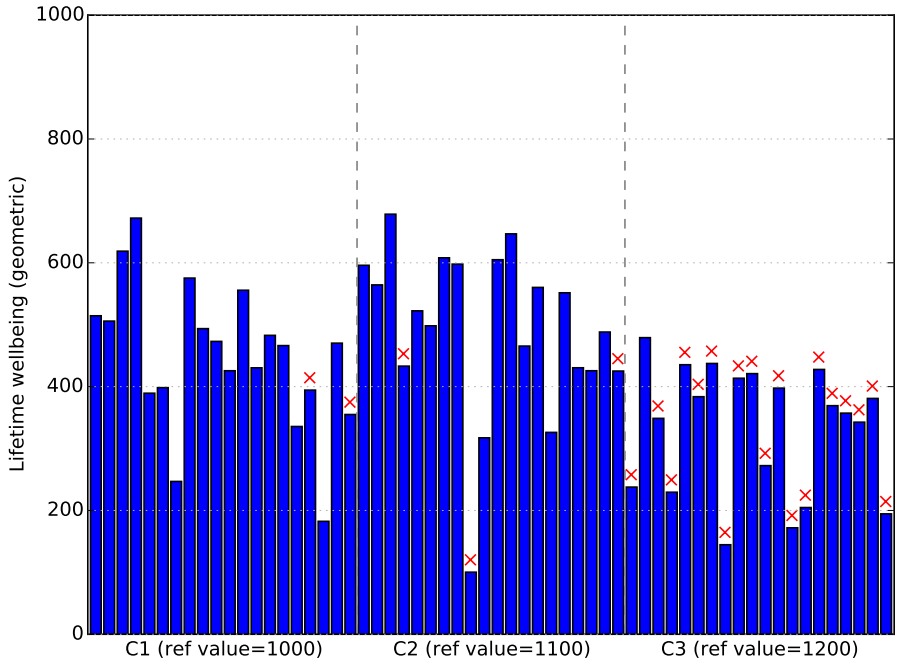

Figure 5. Experimental results. The means of the robot's geometric wellbeing over the lifetime of each run. Larger values indicate better maintained physiological variables. Red crosses indicate runs in which the robot died.

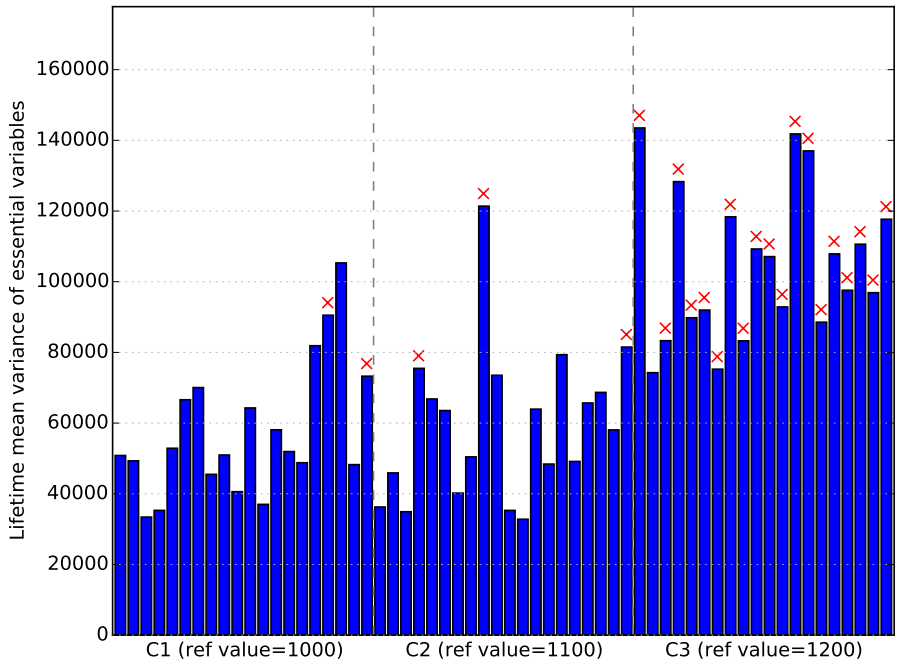

Figure 6. Experimental results. The means of the variance of the robot's physiological variables (which can be thought of as a measure of robot's "physiological balance") over the lifetime of each run. Smaller values indicate better balance between the difference physiological variables. Red crosses indicate runs in which the robot died.

balanced" management of the four physiological variables, although in some scenarios it may be advantageous, e.g. it might be a good strategy for the robot to increase the value of one variable when resources are abundant, if it is likely that the relevant resource will be scarce in the future.

Applying this metric to our experimental data, we calculated the physiological balance in a three-step process. First, in each run in each condition, we calculated the physiological balance at each "time step" (each point in time at which data was collected). Second, we calculated the mean physiological balance over the lifetime of the robot in each run (twenty runs per condition, sixty in total; these values are shown in Figure 6). Finally, we calculated the overall mean for each condition as the mean of the twenty "lifetime means" (one per run) from the previous step. These values are shown in Table 2 (fourth column).

There was a statistically significant difference between conditions (ANOVA, $p < 1 \times 10^{-6}$) with Tukey HSD post-hoc analysis showing that condition 3 (highly unrealistic target values) differed from the other two conditions ($p < 0.001$), although there was no statistically significant difference between conditions 1 and 2.

In summary, as with the wellbeing metric, a highly unrealistic target value is disadvantageous in terms of physiological balance. However, the chosen mildly unrealistic target value (condition 2) is not statistically different from the realistic target value (condition 1) for this metric.

Maintenance of Variables Away from Dangerous Values In order to evaluate how well our robots kept their physiological variables from falling to dangerous values (near to zero, the critical limit of the variables), we calculated the percentage of the robots' lifetime during which any variable was $= 0$, ≤ 100, ≤ 200, ≤ 300. These percentages are shown in Figure 7, and the specific values for a variable $= 0$ in Table 2 (last column). Here, lower values (smaller percentages of time) are better since they indicate less time spent with a variable in the "danger zone" close to the critical limit[7].

In summary, as with the wellbeing metrics, a highly unrealistic target value is disadvantageous in terms of maintaining the physiological variables above the dangerous values. However, contrary to what we would expect, the mildly unrealistic target value is not disadvantageous compared to the realistic target value condition: given our results, it may be advantageous compared to the realistic target value, in that the physiological variables were better maintained away from the low values.

Maintenance of Integument Focusing more closely on the maintenance of the integument variables, we calculated the percentage of the robots' lifetime during which either of the two integument variables were higher than the other two physiological variables ("best maintained") or lower than the other two physiological variables ("worst maintained"). These are shown in Figures 8 and 9. Note that since we have two integuments, one side may be the best maintained, while at the same time the other side is the worst maintained.

[7] In fact, the context plays an important role in determining the significance of the fact that a value is close to its critical limit. For example, if one resource is plentiful, then letting the corresponding physiological variable fall close to its critical limit is not necessarily a bad thing, since it can easily be increased, and it may be more adaptive to try to satisfy the other needs if their related resources are less available.

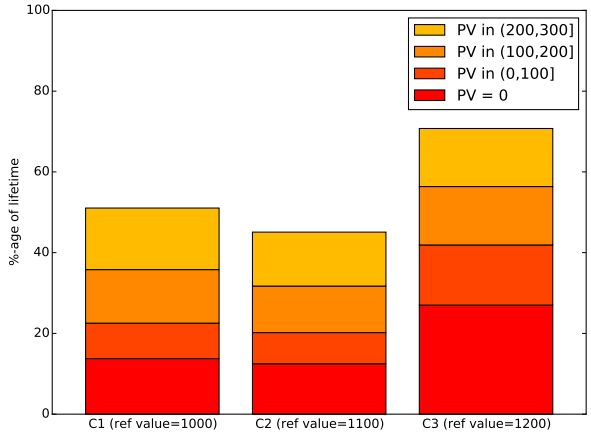

Figure 7. Experimental results. Percentage of the robot's lifetime during which the physiological variable closest to the critical limit of zero was in four "regions" of the physiological space: with a value of exactly 0 (the variable with a value of zero here must be an integument variable, since if it had been one of the survival-related variables, the robot would have been dead), in the range (0,100] (intuitively "highly critical"), in the range (100,200] ("critical"), and in the range (200,300] ("danger"). These percentages were calculated by concatenating the lifetimes of the robots in the twenty runs for each condition, and calculating the percentage of this time during which the physiological variable that was closest to the critical limit was in each region. The equal zero percentages correspond to the values in Table 2, last column.

The chart showing the percentage time as the largest physiological variable shows a trend towards better management of at least one integument. However, this should be considered in the context of the second chart where there is not a clear trend as to the worst managed variable. Looking at the evolution of the physiological variables in condition 3, the robot would often concentrate on grooming one side, at the expense of the other, and the maintenance of integument metrics reflect this fact.

In our robot model, this concentration on grooming one side is connected with the perception of the salience of the grooming post. In our implementation, when grooming is happening, due to the position of the grooming post on one side of the robot, the post cannot be perceived by the IR sensors on the other side of the robot, and therefore the post does not provide an incentive stimulus for the motivation to maintain the integument on the other side of the robot. When the robot has a more realistic (lower) target integument, the action of grooming one side is more likely to be interrupted, providing more opportunities to switch to the other side.

Balance of behaviors In order to evaluate how the robot divided its time among behaviors, we used the internally logged data to calculate the amount of its lifetime spent in grooming and eating, combined over all robots in each condition: Table 2 (fifth and sixth columns). For its remaining lifetime, the robot was searching for a resource (this does not account for all the time spent searching, since occasionally the robot would pass over an energy resource and would eat opportunistically without stopping its search).

As we can see from Table 2, the percentage of time spent grooming increases with the increased perceived ideal value of the integument in conditions 1 to 3. The percentage

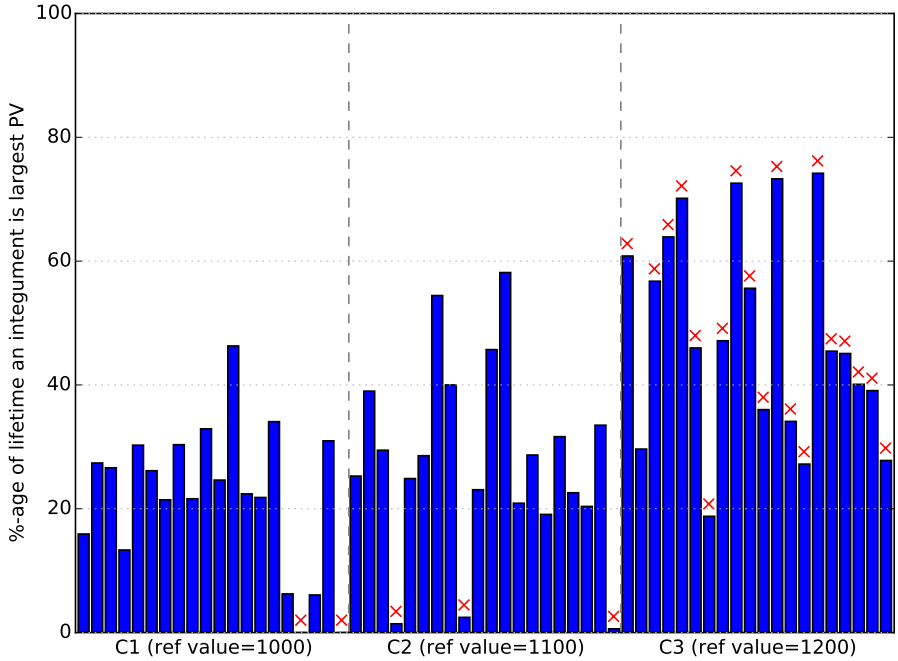

Figure 8. Experimental results. The percentage of the robot's lifetime that either of the two integument variables was the largest valued (i.e. most well maintained) essential variable. Red crosses indicate runs in which the robot died.

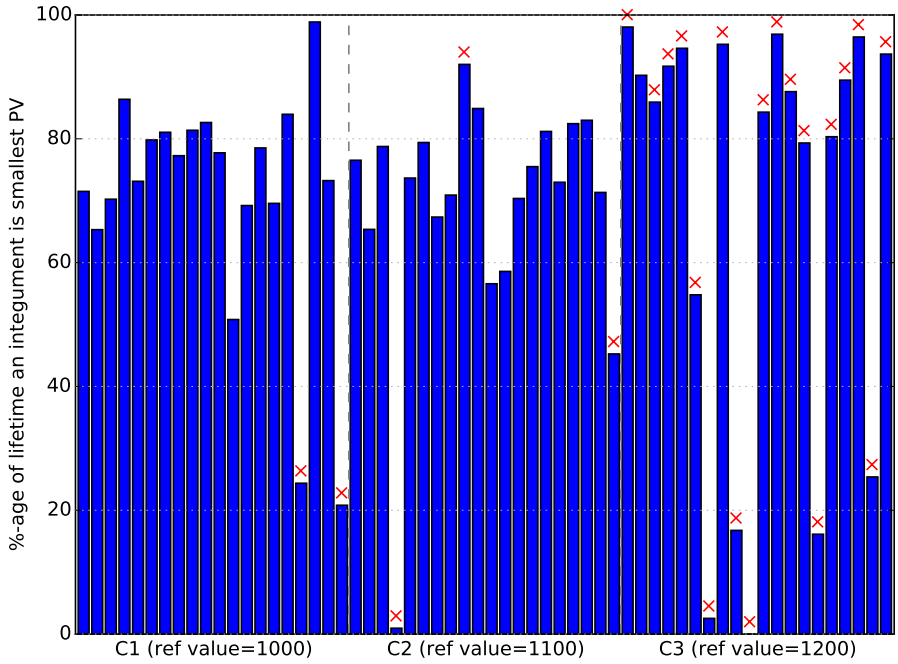

Figure 9. Experimental results. The percentage of the robot's lifetime that either of the two integument variables was the smallest valued (i.e. least well maintained) essential variable (right). Red crosses indicate runs in which the robot died.

Table 3. Experimental results. Percentage of time during which each motivation was the highest, taken as a percentage of the robots' combined lifetime (values in brackets: taken as a percentage of the time when the robot was searching). (Total percentages may exceed 100%, since if two motivational values were equal largest, they were counted in both categories.)

Condition	Feed	Avoid	Groom
1	28.7 (20.5)	6.8 (8.0)	67.5 (78.5)
2	26.4 (18.0)	2.7 (1.8)	71.0 (80.3)
3	13.2 (4.8)	0.32 (0.25)	86.5 (95.2)

time eating decreases as the target integument increases, but not as much as the grooming increases, and there is only a small decrease between conditions 1 and 2. This indicates that the robots with mildly unrealistic target values are spending less time searching, and more time executing consummatory behaviors. This may be an indication that there is some adaptive value in a mildly unrealistic target.

Balance of motivations Since we are considering OC-spectrum disorders, we wanted to check if our robot had anything analogous to obsessions. In order to evaluate the robots' "concerns", we calculated the percentage of the robots' lifetime during which the motivation with the highest intensity was either to feed, avoid or groom. We additionally calculated the corresponding percentages of time during which the robot was wandering around in search of either an energy resource or a grooming post. The results are shown in Table 3.

These figures indicate that all robots spent the majority of their lifetime motivated to groom (i.e. attending to either of the two integuments, either by active grooming or by searching for a grooming post), and the smallest part of their lifetime motivated to avoid objects (since in general the robots did not experience much damage by collisions). As expected, as the perceived integument target value became more unrealistic through the three conditions, the amount of time when the highest motivation was to groom increased, and the amount of time when the other motivations were highest decreased. However, the change was not smooth across the conditions, with a larger change in motivations to feed and to groom occurring from Condition 2 to Condition 3. Although this result is preliminary, it suggests a nonlinear response of the internal motivations ("concerns") to the different perceived ideal values in the different conditions.

Discussion of Experimental Results

Our clearest result is the difference in the number of deaths in the three runs (Table 2, first column). Seventeen of the deaths in condition 3 (highly unrealistic target values) occurred when the robot was grooming, not stopping even though in each case one integument had reached its maximum value, and the energy was falling to its fatal limit. In contrast, the robot in condition 1 (realistic targets) would always stop grooming before the integument reached 1000, enabling it to search for and find an energy resource in time to feed – the only two deaths in condition 1 could both be partially due to "bad luck" with the robot not finding an energy resource while searching for one.

In condition 2 (mildly unrealistic target values), with the more moderate perceptual error, in sixteen of the seventeen runs where the robot survived, at some point in the run,

one or other integument would reach 1000. However, in this condition, it would eventually stop grooming.

In order to examine why grooming continues, we consider the three reasons for the robot to stop grooming:

1. Integument improves sufficiently from grooming, therefore the robot's motivation to groom falls sufficiently that the robot acts to satisfy another motivation.
2. Energy falls so low that the motivation to feed (even in the absence of an energy resource acting as an external cue) exceeds the motivation to groom (even though this is increased by the presence of an external cue).
3. Integrity falls so low that the motivation to avoid objects prompts the robot to move away from the grooming post. This could be due to damage incurred during grooming and interaction with the post.

Our results show that in conditions 2 and 3, the distorted perception interfered with the normal dynamics of decision making, and in particular with the above reasons to stop grooming. In conditions 2 and 3 we see from our data that the first reason to stop grooming was made less likely compared to condition 1, since there were cases where an integument reached its maximum value and the robot did not stop grooming, and it was only the continuing fall in the other variables (increasing the corresponding motivation) that caused the robot to move away. Additionally, in condition 3 (highly unrealistic targets), the second reason to stop grooming was also less likely, since there were cases where the energy fell all the way to zero, and even with maximum integument value, the motivation to groom was still higher than the motivation to feed. In our experimental setup, the rate of damage from grooming was not sufficiently high for the integrity to fall low enough to prompt the robot to move away, and therefore we cannot say if reason 3 was still a factor.

Possible Advantages of an Unattainable Target While the highly unrealistic target values (condition 3) lead to almost inevitable death, examining condition 2 indicates that a mildly unrealistic target value may confer an advantage to our robot. We can see this in several of our results, listed below. With this positive perspective on the unachievable targets, we could characterize them as "idealistic", rather than "unrealistic".

Firstly, an advantage of mildly unrealistic target values can be seen in the increased arithmetic and geometric wellbeings for condition 2 compared to condition 1 (Table 2, second and third columns), indicating better overall "health".

Secondly, if we look at the percentage of lifetime during which our robots had zero (worse maintained) integument (Table 2, last column), we see that in condition 2, the percentage is smaller than for condition 1. In a situation where maintaining integument aids survival (as is typically the case in animals) this smaller amount of time where the value of an integument was zero represents an advantage. We see the same advantage (reduced times for condition 2) in considering the percentage of time that any of the physiological variables were below particular values (Figure 7). The high values for condition 3 reflect that the robot would typically neglect one integument while focusing on the other. However, we are cautious about drawing conclusions from this difference in condition 3, since it may be the shorter lifetimes in condition 3 that make the percentage of lifetime larger.

Thirdly, we can look at the balance between the time spent grooming and the time spent feeding (Table 2, fifth and sixth columns). Comparing condition 2 to condition 1, the percentage of time spent grooming increases by a moderate amount (from 34.5 to 39.4) as expected, but while the time spent feeding does decrease, it does so by only a small amount (from 21.3 to 20.6). This can viewed as due to increased persistence of the consummatory grooming behavior making more use of the resource when it is available. Rather than the time eating, it is principally the time spent searching that is reduced by the increased grooming.

Fourthly, the relatively small penalty that we have observed above for increasing the target value in condition 2 is also apparent in the small increase in variance of the essential variables (Table 2, fourth column, and Figure 6) indicating that the physiological balance is minimally affected.

It should be noted that none of these differences between condition 1 and 2 was found to be statistically significant, which is not unexpected since the value of 1100 was not chosen for the purposes of testing an improved performance in the robot compared to condition 1 (there might be some target value either above or below 1100 where the improved performance is more marked). However, these different lines of evidence all point to advantages of a target value greater than 1000. This question of advantages of a mildly unrealistic value is, therefore, a hypothesis for future research. It may be the case that the optimal value is somewhere between 1000 and 1100, the exact value depending on the metric chosen and the environment, as well as other variables. Such potential advantages of mildly unrealistic target values also contribute to the debate about possible evolutionary origins of OCD (Glass, 2012).

Computed Threshold for Unstoppable Grooming Finally, mathematical analysis of the algorithm used to compute the intensity of the robot's motivations allows us to calculate a theoretical threshold for the perceived target integument value, above which grooming would become unstoppable: a perceived target integument value that results in a grooming behavior that will not be stopped by the motivation to feed. The value is calculated as follows.

First, we need to deduce what the intensity of the motivation to groom is when the grooming behavior continues indefinitely. To do this, let us consider a situation with perfect integument (=1000). In this situation, taking equation 1, in condition 1 the motivation to groom is zero, in condition 2 it ranges from 100 to 600 (depending on the size of the cue), and in condition 3 it can range from 200 to 1200. In this calculation, in which we are considering the general case of an arbitrary target value, our experimental data justifies the use of the maximum value for the cue to groom. Examining our data, we see that, on the occasions when integument reaches its maximum value, the maximum values for motivational intensity are common; hence, the cue in equation 1 must also be at its maximum value.

In considering the competition between the motivations to groom and to feed whilst grooming is ongoing, we will assume that there is no energy resource detected (this assumption is realistic in our environment due to the separation of the resources and the short range of the robot's sensors). Therefore, by equation 1, the intensity of the motivation to feed is equal to the energy deficit, and thus reaches a maximum value of 1000. Hence, we seek the target value T for integument that would give a motivational intensity greater than

1000 for a maximum cue (100) and perfect integument (1000). Substituting from equation 1:

$$motivation_{groom} \geq motivation_{feed}$$
$$(T - 1000) + 100 \times \alpha \times (T - 1000) \geq 1000$$

Solving for T with our choice of $\alpha = 0.05$, we get a minimum value of $T = 1166.7$, and for target integument values above this, our robot will be very unlikely to stop grooming once started. This is in agreement with our experimental results, where in condition 3 our perceived target value (1200) is greater than T, and the robot was highly likely to die from lack of energy while grooming.

Larger values of α would result in values of T closer to 1000; therefore, from this calculation, we can predict that smaller errors in the perceived target value would result in pathological behavior.

ROBOT MODEL DESIGN PROCESS: POST-DEVELOPMENT STAGES

Evaluation of the Robot Model (Stage 7)

After assessing the model through analysis of the experimental results, we now expose the robot model and its underlying theoretical model to criticism in order to evaluate the quality of the model and, in subsequent stages, whether it is of clinical use and how to improve it in the next iteration of our design process.

In order to evaluate our robot and its interaction as a model of an OC-spectrum disorder, we consider four criteria based on their use to evaluate animal models: *face validity*, *construct validity*, *predictive validity* and *reliability* (Geyer & Markou, 2000; Lewis & Cañamero, 2017; van der Staay, 2006).

Face Validity Face validity refers to the descriptive similarity (van der Staay et al., 2009) or "phenomenological" similarity (Willner, 1986) between the robot model and specific features of the phenomenon that is being modeled, in this case, OC-spectrum disorders. This similarity would concern, for example, a specific symptom or a behavioral dysfunction observed in both the patient and the robot model, and is not related to the experiential quality that the term "phenomenological" has in philosophy (in the phenomenological tradition). Therefore, the robot behavior should resemble the OC-spectrum disorders being modeled by showing features of the disorders, and not showing features that are not seen in the disorders.

Our results show that we achieved high face validity within the scope of our model, focused on compulsions and obsessions: the self-grooming behavior was executed for long periods in conditions 2 and 3, and itself is related to OC-spectrum conditions TTM and PSP. The continuation of the grooming behavior beyond the point where the condition 1 robot would have stopped, can be viewed as perfectionism – a characteristic of several OC-spectrum and related disorders (OCD, TTD, body dysmorphic disorder, OCPD) (Fineberg et al., 2015; Fineberg, Sharma, Sivakumaran, Sahakian, & Chamberlain, 2007; Pélissier & O'Connor, 2004). Hence, our work provides experimental support for the theoretical claims about Pitman's model being able to generate persistent repetitive behavior – i.e. that Pitman's model can generate behavior with face validity.

Our model also shows face validity with respect to the sense of "incompleteness" – an inner sense of imperfection, or the perception that actions or motivations have been incompletely achieved (Hellriegel, Barber, Wikramanayake, A. Fineberg, & Mandy, 2016; Pitman, 1987; Summerfeldt, 2004; Wahl, Salkovskis, & Cotter, 2008) – which is widely viewed as a key aspect of OCD and can be linked with our persistent internally sensed error. In our motivation-based architecture, the motivational systems are goal-oriented embodied sensorimotor loops, and in the pathological case the perceived need is never satiated, even if an outside observer would say that the goal – grooming to improve integument – has been achieved. In other words, in the pathological case, goal-oriented behavior is never complete because the perceived need is never satiated, and therefore the corrective behavior continues even if the error of the physiological variable has actually been corrected.

In addition, considering the results from our experiments regarding maintenance of integument, the concentration of the robot on grooming one side, even to the neglect of the other side, bears a potential phenomenological similarity with PSP, in which, in some cases, a person may concentrate their skin picking in one place, causing skin lesions.

Our model does not yet include other characteristics of OCD, such as additional non-functional ritual behaviors (Amitai et al., 2017; Eilam, Zor, Fineberg, & Hermesh, 2012) or indecision aspects (Sachdev & Malhi, 2005) that can occur in OCD and TTM. At its present level of development, our model lacks some key mechanisms hypothesized to be behind such non-functional ritual behaviors. For example, our robot has no learning capability, and therefore if the development of non-functional rituals is, as some theorize (Eilam, 2017), due to a disrupted behavior learning process, there is no opportunity the robot to develop such learned rituals. Similarly, indecision is theorized as resulting from an inability to choose between strong competing goals (Pitman, 1987). However, our current model has limited capacity for such conflict, because in the experimental setup presented here, only one resource (and hence only one cue for action) would be detectable by the robot's short-range sensors, so such competition would be unlikely.

Adding further complexity would allow us to produce such non-functional ritual behaviors using different mechanisms such as those mentioned above, and to compare experimentally these different hypotheses.

In summary, although our robot's behavior does not exactly match an OC-spectrum condition, it exhibits those aspects that we would expect, given the scope and complexity of our model. A simple model like ours also allows for an incremental investigation of the behavior, in which different aspects of the condition are the result of specific additions to the model.

Construct Validity Construct validity indicates the degree of similarity between the underlying mechanisms of the model and of the condition (Epstein, Preston, Stewart, & Shaham, 2006). In the context of animal models, Joel (2006) specifies that the underlying mechanisms for construct validity may be either physiological or psychological. According to van der Staay et al. (2009), construct validity reflects the "soundness of the theoretical rationale". In our robot model, we don't directly model psychological constructs, and we see them as being more related to face validity (e.g. the sense of incompleteness discussed in the previous section).

When talking about underlying mechanisms, animal models have tended to focus on specific elements, such as the involvement of specific brain areas, receptors, chemicals or genes (Camilla d'Angelo et al., 2014, Table 1). Construct validity in this case is then based on finding specific mechanisms underlying a phenomenon (e.g. symptoms) in the animal model and the human condition. Such a view of construct validity might be critically questioned by approaches that emphasize species-specific features and differences, such as models grounded in ethology.

Such a view of construct validity also implies that the idea of robot models having construct validity might be problematic, and questioned on various 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 allows us to capture key selected structural, functional, or dynamics elements for a focused, rigorous investigation. The adoption of a cybernetics perspective that focuses on interaction dynamics, processes and general principles, also means that we can model aspects of underlying physiological mechanisms relevant to OCD that are more general than the specific types of underlying mechanism that animal models have focused on; for example, we can implement processes that model effects on perception that may be hypothesized to involve specific chemicals, without having to model the specific chemicals themselves. In addition, in robot models, mechanisms underlying a phenomenon can be modeled at different levels of granularity from different theoretical perspectives. These complementary constructs, models and levels could be experimentally tested and compared, bridging gaps across levels and conceptual perspectives, which is a crucial issue in cross-disciplinary and translational research.

From a conceptual perspective, construct validity for our robot model would be linked to the construct validity of the cybernetic and signal attenuation models of OCD, since the underlying mechanisms that we use to model OCD are closely related to the mechanism they postulate. In all of the three models (the robot model and the two conceptual models), the emphasis is on the dynamics of interaction among the elements of a regulatory system, rather than than attempting to locate the problem in and modeling specific brain areas or specific genes. All these models share the use of cybernetics notions, and conceive of OCD as a disorder in the decision making process, in particular the presence of a high error signal that cannot be eliminated by behavioral output. One of the causes that Pitman proposes for this persistent high error signal is an intrinsic comparator defect, and our pathological case is generated by a fault in the robot's comparator system that gives rise to an error signal that cannot be eliminated through the robot's behavior. In terms of the signal attenuation model, the robot's behavior does not result in feedback as to the success of the behavior since the error signal remains high, so the behavior continues.

Whereas the signal attenuation model has received more attention and has provided more examples of construct validity, we have found little direct investigation of Pitman's model as it applies to humans with OC-spectrum disorders; it thus remains largely theoretical. At this early stage of our research, we can therefore only claim limited and indirect construct validity for our robot model.

Robot models that we have used in previous work, based on similar motivational architectures, have included elements that could allow us to link to anxiety, perception of harm or an excessive reliance on habits (as opposed to instrumental acts), all of which have

been the basis of conceptual models of OCD. This could potentially allow us to expand the construct validity of our model in relation to other theoretical models.

Reliability An animal or robot model is said to be reliable if the experimental outputs are reproducible (in the sense that the exact experiments can be reproduced, possibly by different experimenters, producing the same results) and extended replications can be run (e.g. conceptual replications, in which the same underlying concepts can be tested in different ways).

As a robot model with an explicitly programmed controller and a highly controlled environment, we would expect the reliability of our model to be high (highly reproducible) compared to animal models. Indeed, with relatively few runs, we obtained statistically significant results.

Predictive Validity Predictive validity indicates that the behavior of the robot model can be used to make reliable predictions about the condition being modeled. A particularly important aspect of this is that the model can be used to make predictions about outcomes in the human condition, and which interventions will, or will not, work with some degree of accuracy.

This aspect of predictive validity, highly important for clinical research purposes, is currently lacking in our model. In our experiments we did not investigate any "treatments", and at this early stage in the development of the model, we could expect only limited predictive validity. However, this is an important point to be developed in future work, so that our experiments can inform future clinical research.

Is the Robot Model Sufficiently Advanced to be of Clinical Use? (Decision Point 8)

We consider if our model might be sufficiently advanced for clinical use. If the model was sufficiently advanced, then we would move to Stage 10 in our design process; otherwise, we need to ask if the results so far indicate if there is a potential for improvement or not (see Decision Point 9 below).

Strictly speaking, at this point, the answer is that our model is not yet sufficiently advanced for clinical use. However, we already have suggestions for potential avenues to clinical use. Even at the current stage of development, the robot could be used as a working model to help OCD patients understand their condition, and reduce negative feelings about it. For example, seeing how compulsive behavior in the robot results from the perception of a high persistent error that is not corrected through behavior, can help them understand, and feel relieved, that similar behavior in them may be the result of a processing error, rather than their often held assumption that they are "morally wrong", which can be very emotionally disturbing for them. Although, to our knowledge, robots have not been used in the treatment of OC-spectrum disorders, they have been used as therapeutic tools in other areas, for example in autism spectrum disorder (ASD) (see Diehl, Schmitt, Villano, and Crowell (2012), Pennisi et al. (2016) for reviews). However, our proposed use would differ significantly from this other robots, since they are tools to be used in therapy, mostly as stimuli for interaction, but they do are not models of the condition (they do not "have" the condition) whereas our robot is a model (it "has" an OC-spectrum disorder) that we also aim to use as a tool. A closer match for our proposed use would be to our own robot Robin, which is controlled by a related software architecture and includes a model of diabetes.

Robin was designed as a tool to support diabetes education and management in children with diabetes (L. Cañamero & Lewis, 2016), focusing particularly on affective elements of diabetes self-management (Lewis & Cañamero, 2014).

Do the results so far indicate the potential for improvement? (Decision Point 9)

At this stage, we need to ask if the results so far indicate if there is a potential for improvement, particularly with respect to our evaluation criteria, and in the direction of clinical relevance. Let us thus assess potential improvements to our model in the direction of clinical applications.

One of the main treatments for OCD is exposure and response prevention (ERP), which involves habituation to the urges to perform compulsions, resulting in the compulsions being extinguished (Storch & Merlo, 2006). Currently in our model, while we could prevent the robot from grooming, there is no adaptive capability in its controller that would make this change its future behavior. Both this and our model's lack of capability necessary to develop non-functional rituals point to a direction for future research: introducing adaptation in its behavior. This could be done, for example, by making reference values susceptible to change (modulation) through external environmental factors, such as exposure "treatments"; by adding receptors for the internal signals that could in turn be modulated by long-term signal strength, allowing reinforcement or habituation of behaviors (Lones, Lewis, & Cañamero, 2018); or by adding a capability to inhibit behaviors thus separating obsessions and compulsions. Such additions would be aimed at improving the model's face validity (non-functional behaviors) and predictive validity (potential treatments) and hence improving its clinical relevance.

Allowing the robot to adapt and respond to treatment in this way may also provide another avenue to clinical application. Showing the working robot model to patients, as in the previous section, but in this case, also showing the patient the robot's improvement after applying therapy to the robot, might help them to understand and accept the often stressful ERP treatment, in which they are exposed to the triggers for their compulsions.

Accept the robot model for use in clinical studies (Stage 10)

In the case that the robot model was sufficiently advanced for clinical trials (Decision Point 8 above) we would move the Stage 10: "Accept the (improved) robot model for use in clinical studies". This contrasts with the process so far, which has concentrated on model development, more in the domain of robotics research. The details of this stage would depend on the proposed clinical research. One possible route would be to investigate potential treatments by manipulating targeted elements of the model in different ways, either internally in the robot (e.g. by amplifying particular internal signals where problems have been hypothesized in humans), or externally, in the environment (e.g. by exposing the robot to problem situations in order to analyze whether and how it adapts). If adjusting an element of the robot model reduces symptoms in the robot, then the analogous adjustment in human patients could be investigated as potential targets for intervention. We expect that, initially, such applications of the robot model would result in very broad targets for intervention, but as the robot model is refined, these predictions could also be refined.

In any case, even as more clinically-focused research begins, development of the robot model would continue, following the process described in this paper. However, feedback from the clinical researchers could be brought to different stages of the robot model devel-

opment process. For example, phenotypical targets in the selection stage (Stage 3) could be drawn from observations in human subjects in the clinical studies, or from elements that are theorized as potential pharmaceutical targets. As another example, the design of robot experiments (Stage 6) could be done with the design of corresponding clinical studies in mind.

Is further refinement of the robot model required? (Decision Point 11)

This decision point is similar to Decision Point 9 above, with the difference that, in point 9, the robot model has not yet been considered sufficiently advanced for clinical research. Consequently, if further model development was not possible, then the robot model would be rejected as inadequate for clinical use, although it may still shed light on the underlying theoretical model used as the basis for the robot model. In contrast, at Decision Point 11, the model is already considered sufficiently advanced for some clinical research, and this research can potentially continue even as model development stops.

Induction Stage (Stage 12)

Having given some indication of how we can refine our model, we now reach the Induction stage. Here we use the knowledge gained from the Evaluation stage to refine our assumptions and definitions, both those identified at the Consensus stage, and any implicit assumptions that we had made and not identified. In our case, we see that we should think carefully about the different properties of what we have called variously a "target", "reference" or perceived "ideal" value, and the generation of the error signal, and what the range of adaptive values might be. Specifically, a "good" reference value for the comparator mechanism for a cybernetic model may not be one that is achievable, and an error signal that can never be reduced to zero may not be an indication of a pathology.

While in our case, the Induction stage has shed light on an underlying assumption of the cybernetic model, and hence is relevant to research into both conceptual and robot models, the induction stage may also re-evaluate the assumptions made about the clinical aspects of the model. For example, the nature of the phenotypes might be reconsidered if the behavior of the robot deviated in some unexpected way from the clinical description, perhaps by showing additional behaviors or internal states. These unexpected observations could indicate either that the model was in error, or that the clinical description of the condition was incomplete.

CONCLUSIONS

In this paper, we have discussed and illustrated the use of robot models to complement existing computational and animal models in psychiatric research. We have described a design process for robot models of mental disorders stemming from animal models, and illustrated this design process with the initial development of a robot model for OC-spectrum disorders, including initial experiments and results. Our model builds on our work on architectures for decision making in autonomous robots, and also on existing models of OCD – specifically the cybernetic model and the signal attenuation model – to link with existing research. The design process has also given directions for future work with a view to the model's clinical relevance.

Although this initial stage of development only models the most basic aspects of such disorders, and does not approach the complexity of OCD in humans, our results already

serve to shed light on aspects of the theoretical model on which they are based that are not obvious simply from consideration of the model: specifically the non-linear relationship between the perceived target value and the onset of pathological behavior, and the possible advantage of a mildly unrealistic target. This result might have implications in clinical research and treatment, for example by helping us understand why some members of a family develop OCD while others do not.

This initial development work on a robot model has also generated a hypothesis for future research: that mildly unrealistic target values may provide some advantages for our robot. Such potential advantages may also be explored in humans, in animal models and in cybernetic systems in general.

To conclude, we would like to add some remarks on the nature of robot models and their relation to other models relevant to computational psychiatry.

As models, robots present very different features to other types of models such as computational models or simulated environments. To characterize the main differences between computational and robot models, we find it useful to think of the distinction that Herbert Simon, one of the founders of Artificial Intelligence, drew between types of models in his book "The Sciences of the Artificial" (Simon, 1981), when trying to characterize the meaning of the terms "artificial" and "simulation". Simon distinguished between models that simulate a system by predicting its behavior and deriving consequences from premises (e.g., a system for weather prediction), and models that are a simulation of a system by embodying a few key features of that system and being put to behave in the same environment, governed by the same laws (e.g., a satellite is not a simulation of a moon, it is a moon, the "real thing"). While computational models fall in the first category, embodied autonomous robot models, such as ours, fall in the second. According to Simon, the first type of models are appropriate for achieving understanding of systems with many parameters, for which it is difficult to predict behavior without complex or extensive calculations, whereas the second type is most useful as a source of new knowledge to understand, by synthesis, the behavior of poorly understood systems. The choice between one or the other type of model will depend on the type of research questions under investigation.

Some would perhaps argue that a simulated agent in a simulated environment might also belong to the second type of models and might be preferable to robot situated in the physical world because replicability of experiments can be higher. We do not think such type of models belong to the second category, but to the first. The complexity (including important features such as unpredictability and "noise") of physical world, a physical agent, and their interactions, cannot be fully simulated (Brooks, 1991b; Pfeifer & Scheier, 2001). In a simulated environment, we can only see the consequences of the features that we have included in it, even if we simulate some noise and unpredictability; however, in the real world, unexpected noise and unpredictable elements that we had not anticipated might give rise to significant behavior. This is the case in both robots and humans. As a "trade-off", these features might reduce exact replicability, although replicability is still very high when using robots and, if data are properly logged during experiments, it is often possible to analyze when unexpected behavior might be due to noise. In the other direction, this "trade-off" means that the easier replicability of experiments using a simulated agent in a simulated environment comes at the cost of an impoverished model that might leave out features that had not been anticipated by the designer, but might end up being significant. Therefore, in addition to the same considerations made above regarding the two

different types of simulations distinguished by Simon, the choice between a physical robot situated in the physical (and social) environment, and a simulated agent in a simulated environment, also depends on how important features such as dynamics of interaction or embodied sensorimotor loops, are to address the question under investigation.

ACKNOWLEDGMENTS

The authors thank David Wellsted of the University of Hertfordshire, and Valerie Voon of the University of Cambridge for fruitful discussions, and two anonymous reviewers for their comments and suggestions to improve the quality and clarity of the manuscript. ML is supported by an Early Career Research Fellowship grant from the University of Hertfordshire, awarded to LC.

AUTHOR CONTRIBUTIONS

The project was formulated by LC. The model in this paper was conceptualized and designed equally by LC and ML, and implemented in software by ML with support from LC. Experiments were designed by ML and LC, with clinical advice from NF. They were run and analyzed by ML, with support from LC and NF. Writing of the initial draft was lead by ML, with support from LC and NF. Editorial was mostly done by LC. Reviews were addressed by ML and LC, with input from NF.

REFERENCES

Adams, R. A., Huys, Q. J. M., & Roiser, J. P. (2016). Computational psychiatry: towards a mathematically informed understanding of mental illness. *Journal of Neurology, Neurosurgery & Psychiatry*, *87*(1), 53–63. doi: 10.1136/jnnp-2015-310737

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (5th ed.). Arlington, VA: American Psychiatric Association.

Amitai, M., Arnon, N., Shaham, N., Gur, S., Apter, A., Weizman, A., & Hermesh, H. (2017). Nonfunctional redundant acts characterize OCD, even in OCD-unrelated tasks: A demonstration in questionnaire completion. *Psychopathology*, *50*(6), 389-400. doi: 10.1159/000479885

Apergis-Schoute, A. M., Gillan, C. M., Fineberg, N. A., Fernandez-Egea, E., Sahakian, B. J., & Robbins, T. W. (2017). Neural basis of impaired safety signaling in Obsessive Compulsive Disorder. *Proceedings of the National Academy of Sciences*, *114*(12), 3216–3221. doi: 10.1073/pnas.1609194114

Ashby, W. R. (1960). *Design for a brain; the origin of adaptive behavior* (2nd ed.). New York, NY: John Wiley & Sons.

Avila-García, O., & Cañamero, L. (2004). Using hormonal feedback to modulate action selection in a competitive scenario. In S. Schaal, A. J. Ijspeert, A. Billard, S. Vijayakumar, J. Hallam, & J.-A. Meyer (Eds.), *From animals to animats 8: Proc. 8th intl. conf. on simulation of adaptive behavior (SAB'04)* (pp. 243–252). Los Angeles, USA: MIT Press.

Brooks, R. A. (1991a). Intelligence withour reason. In *Proc. 12th international joint conference on artificial intelligence (ijcai'91)* (pp. 1–27). Sydney, Australia.

Brooks, R. A. (1991b). New approaches to robotics. *Science*, *253*(5025), 1227–1232. doi: 10.1126/science.253.5025.1227

Camilla d'Angelo, L.-S., Eagle, D. M., Grant, J. E., Fineberg, N. A., Robbins, T. W., & Chamberlain, S. R. (2014). Animal models of obsessive-compulsive spectrum disorders. *CNS Spectrums*, *19*(1), 28–49. doi: 10.1017/S1092852913000564

Cañamero, L., & Avila-García, O. (2007). A bottom-up investigation of emotional modulation in competitive scenarios. In A. C. R. Paiva, R. Prada, & R. W. Picard (Eds.), *Proc. second international conference on affective computing and intelligent interaction (ACII 2007)* (Vol. 4738, pp. 398–409). Lisbon, Portugal: Springer Berlin Heidelberg. doi: 10.1007/978-3-540-74889-2_35

Cañamero, L., & Lewis, M. (2016). Making new "new ai" friends: Designing a social robot for diabetic children from an embodied ai perspective. *International Journal of Social Robotics*, *8*, 523–537. Retrieved from http://dx.doi.org/10.1007/s12369-016-0364-9 doi: 10.1007/s12369-016-0364-9

Cañamero, L. D. (1997). Modeling motivations and emotions as a basis for intelligent behavior. In W. L. Johnson (Ed.), *Proceedings of the first international symposium on autonomous agents (Agents'97)* (pp. 148–155). Marina del Rey, CA, USA: The ACM Press. doi: 10.1145/267658.267688

Colgan, P. W. (1989). Ethology of motivation. In *Animal motivation*. Dordrecht: Springer.

Corlett, P. R., & Fletcher, P. C. (2014). Computational psychiatry: a Rosetta Stone linking the brain to mental illness. *The Lancet Psychiatry*, *1*(5), 399–402. doi: 10.1016/S2215-0366(14)70298-6

Damasio, A. (2010). *Self comes to mind: Constructing the conscious brain*. London: William Heinemann.

Diehl, J. J., Schmitt, L. M., Villano, M., & Crowell, C. R. (2012). The clinical use of robots for individuals with autism spectrum disorders: A critical review. *Research in Autism Spectrum Disorders*, *6*(1), 249–262. doi: 10.1016/j.rasd.2011.05.006

Eilam, D. (2017). From an animal model to human patients: An example of a translational study on obsessive compulsive disorder (OCD). *Neuroscience & Biobehavioral Reviews*, *76*, 67–76. doi: 10.1016/j.neubiorev.2016.12.034

Eilam, D., Zor, R., Fineberg, N., & Hermesh, H. (2012). Animal behavior as a conceptual framework for the study of obsessive–compulsive disorder (OCD). *Behavioural Brain Research*, *231*(2), 289–296. (Quo Vadis Behavioral Neuroscience: A Festschrift for Philip Teitelbaum) doi: 10.1016/j.bbr.2011.06.033

Epstein, D. H., Preston, K. L., Stewart, J., & Shaham, Y. (2006). Toward a model of drug relapse: an assessment of the validity of the reinstatement procedure. *Psychopharmacology*, *189*(1), 1–16. doi: 10.1007/s00213-006-0529-6

Fineberg, N. A., Apergis-Schoute, A. M., Vaghi, M. M., Banca, P., Gillan, C. M., Voon, V., … Robbins, T. W. (2018). Mapping compulsivity in the DSM-5 obsessive compulsive and related disorders: Cognitive domains, neural circuitry, and treatment. *International Journal of Neuropsychopharmacology*, *21*(1), 42–58. doi: 10.1093/ijnp/pyx088

Fineberg, N. A., Chamberlain, S. R., Hollander, E., Boulougouris, V., & Robbins, T. W. (2011). Translational approaches to obsessive-compulsive disorder: From animal models to clinical treatment. *British Journal of Pharmacology*, *164*(4), 1044–1061. doi: 10.1111/j.1476-5381.2011.01422.x

Fineberg, N. A., Day, G. A., de Koenigswarter, N., Reghunandanan, S., Kolli, S., Jefferies-Sewell, K., … Laws, K. R. (2015). The neuropsychology of obsessive-compulsive personality disorder: a new analysis. *CNS Spectrums*, *20*(5), 490–499. doi: 10.1017/S1092852914000662

Fineberg, N. A., Reghunandanan, S., Kolli, S., & Atmaca, M. (2014). Obsessive-compulsive (anankastic) personality disorder: Toward the ICD-11 classification. *Revista Brasileira de Psiquiatria*, *36*, 40–50. doi: 10.1590/1516-4446-2013-1282

Fineberg, N. A., Sharma, P., Sivakumaran, T., Sahakian, B., & Chamberlain, S. (2007). Does obsessive-compulsive personality disorder belong within the obsessive-compulsive spectrum? *CNS Spectrums*, *12*(6), 467–482. doi: 10.1017/S1092852900015340

Fish, F., Casey, P., & Kelly, B. (2008). *Fish's clinical psychopathology: Signs and symptoms in psychiatry* (3rd ed.). London, UK: Gaskell.

Frijda, N. H. (1986). *The emotions*. Cambridge, U.K.: Cambridge University Press.

Geyer, M. A., & Marcou, A. (2002). The role of preclinical models in the development of psychotropic drugs. In K. L. Davis, J. T. Coyle, & C. Nemeroff (Eds.), *Neuropsychopharmacology: The fifth generation of progress* (pp. 445–455). Philadelphia, PA: Lippincott Williams & Wilkins.

Geyer, M. A., & Markou, A. (2000). Animal models of psychiatric disorders. In F. E. Bloom & D. J. Kupfer (Eds.), *Psychopharmacology: The fourth generation of progress* (pp. 787–798). Philadelphia, PA: Lippincott Williams & Wilkins.

Gillan, C. M., Papmeyer, M., Morein-Zamir, S., Sahakian, B. J., Fineberg, N. A., Robbins, T. W., & de Wit, S. (2011). Disruption in the balance between goal-directed behavior and habit learning in obsessive-compulsive disorder. *American Journal of Psychiatry*, *168*(7), 718–726. doi: 10.1176/appi.ajp.2011.10071062

Gillan, C. M., & Robbins, T. W. (2014). Goal-directed learning and obsessive–compulsive disorder. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *369*(1655). doi: 10.1098/rstb.2013.0475

Glass, D. J. (2012). Evolutionary clinical psychology, broadly construed: Perspectives on obsessive-compulsive disorder. *Evolutionary Behavioral Sciences*, *6*(3), 292–308. doi: 10.1037/h0099250

Hellriegel, J., Barber, C., Wikramanayake, M., A. Fineberg, N., & Mandy, W. (2016). Is "not just right experience" (NJRE) in obsessive-compulsive disorder part of an autistic phenotype? *CNS Spectrums*, *22*(1), 1–10. doi: 10.1017/S1092852916000511

Hezel, D. M., Riemann, B. C., & McNally, R. J. (2012). Emotional distress and pain tolerance in obsessive-compulsive disorder. *Journal of Behavior Therapy and Experimental Psychiatry*, *43*(4), 981–987. doi: 10.1016/j.jbtep.2012.03.005

Huys, Q. J. M., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, *19*(3), 404–413. doi: 10.1038/nn.4238

Joel, D. (2006). Current animal models of obsessive compulsive disorder: A critical review. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *30*(3), 374–388. doi: 10.1016/j.pnpbp.2005.11.006

Lehner, P. N. (1996). *Handbook of ethological methods* (2nd ed.). Cambridge, U.K.: Cambridge University Press.

Lewis, M., & Cañamero, L. (2014). An affective autonomous robot toddler to support the development of self-efficacy in diabetic children. In *Proc. 23rd annual IEEE international symposium on robot and human interactive communication (IEEE RO-MAN 2014)* (pp. 359–364). Edinburgh: IEEE. doi: 10.1109/ROMAN.2014.6926279

Lewis, M., & Cañamero, L. (2016). Hedonic quality or reward? a study of basic pleasure in homeostasis and decision making of a motivated autonomous robot. *Adaptive Behavior*, *24*, 267–291. doi: 10.1177/1059712316666331

Lewis, M., & Cañamero, L. (2017). Robot models of mental disorders. In *Proc. 7th international conference on affective computing and intelligent interaction, workshops and demos (ACIIW 2017)* (pp. 193–200). San Antonio, TX: IEEE. doi: 10.1109/ACIIW.2017.8272613

Lones, J., Lewis, M., & Cañamero, L. (2018). A hormone-driven epigenetic mechanism for adaptation in autonomous robots. *IEEE Transactions on Cognitive and Developmental Systems*, *10*, 445–454. doi: 10.1109/TCDS.2017.2775620

Maia, T. V., & Cano-Colino, M. (2015). The role of serotonin in orbitofrontal function and obsessive-compulsive disorder. *Clinical Psychological Science*, *3*(3), 460–482. doi: 10.1177/2167702614566809

Mantz, S. C., & Abbott, M. J. (2017). The relationship between responsibility beliefs and symptoms and processes in obsessive compulsive disorder: A systematic review. *Journal of Obsessive-Compulsive and Related Disorders*, *14*, 13–26. doi: 10.1016/j.jocrd.2017.04.002

Meyer, V. (1966). Modification of expectations in cases with obsessional rituals. *Behaviour Research and Therapy*, *4*(4), 273–280.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, *16*(1), 72–80. doi: 10.1016/j.tics.2011.11.018

Panksepp, J. (1998). *Affective neuroscience*. New York, NY: Oxford University Press.

Pélissier, M.-C., & O'Connor, K. (2004). Cognitive-behavioral treatment of trichotillomania, targeting perfectionism. *Clinical Case Studies*, *3*(1), 57–69. doi: 10.1177/1534650103258973

Pennisi, P., Tonacci, A., Tartarisco, G., Billeci, L., Ruta, L., Gangemi, S., & Pioggia, G. (2016). Autism and social robotics: A systematic review. *Autism Research*, *9*(2), 165–183. doi: 10.1002/aur.1527

Pessoa, L. (2013). *The cognitive-emotional brain*. Cambridge, MA: MIT Press.

Pfeifer, R., & Scheier, C. (2001). *Understanding intelligence*. Cambridge, MA: MIT Press.

Pitman, R. K. (1987). A cybernetic model of obsessive-compulsive psychopathology. *Comprehensive Psychiatry*, *28*(4), 334–343. doi: 10.1016/0010-440X(87)90070-8

Powers, W. T. (1973). *Behavior: The control of perception*. Aldine.

Sachdev, P. S., & Malhi, G. S. (2005). Obsessive–compulsive behaviour: A disorder of decision-making. *Australian & New Zealand Journal of Psychiatry*, *39*(9), 757–763. doi: 10.1080/j.1440-1614.2005.01680.x

Salkovskis, P., Wroe, A., Gledhill, A., Morrison, N., Forrester, E., Richards, C., … Thorpe, S. (2000). Responsibility attitudes and interpretations are characteristic of obsessive compulsive disorder. *Behaviour Research and Therapy*, *38*(4), 347–372. doi: 10.1016/S0005-7967(99)00071-6

Sergeant, J. (2000). The cognitive-energetic model: an empirical approach to Attention-Deficit Hyperactivity Disorder. *Neuroscience & Biobehavioral Reviews*, *24*(1), 7–12. doi: 10.1016/S0149-7634(99)00060-3

Shafran, R. (2005). Cognitive-behavioral models of OCD. In J. S. Abramowitz & A. C. Houts (Eds.), *Concepts and controversies in obsessive-compulsive disorder* (pp. 229–260). Boston, MA: Springer. doi: 10.1007/0-387-23370-9_13

Simon, H. A. (1981). *The sciences of the artificial* (2nd ed.). Cambridge, MA: MIT press.

Spier, E., & McFarland, D. (1997). Possibly optimal decision-making under self-sufficiency and autonomy. *Journal of theoretical biology*, *189*(3), 317–331.

Stephan, K. E., & Mathys, C. (2014). Computational approaches to psychiatry. *Current Opinion in Neurobiology*, *25*, 85–92. doi: 10.1016/j.conb.2013.12.007

Storch, E. A., & Merlo, L. J. (2006). Obsessive-compulsive disorder: Strategies for using CBT and pharmacotherapy. *Journal of Family Practice*, *55*(4), 329–334.

Summerfeldt, L. J. (2004). Understanding and treating incompleteness in obsessive-compulsive disorder. *Journal of Clinical Psychology*, *60*(11), 1155–1168. doi: 10.1002/jclp.20080

Tyrrell, T. (1993). *Computational mechanisms for action selection* (Doctoral dissertation). University of Edinburgh.

van der Staay, F. J. (2006). Animal models of behavioral dysfunctions: Basic concepts and classifications, and an evaluation strategy. *Brain Research Reviews*, *52*(1), 131–159. doi: 10.1016/j.brainresrev.2006.01.006

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

Verbruggen, F., & Logan, G. D. (2008). Response inhibition in the stop-signal paradigm. *Trends in Cognitive Sciences*, *12*(11), 418–424. doi: 10.1016/j.tics.2008.07.005

Wahl, K., Salkovskis, P. M., & Cotter, I. (2008). 'I wash until it feels right': The phenomenology of stopping criteria in obsessive–compulsive washing. *Journal of Anxiety Disorders*, *22*(2), 143–161. doi: 10.1016/j.janxdis.2007.02.009

Wang, X.-J., & Krystal, J. H. (2014). Computational psychiatry. *Neuron*, *84*(3), 638–654. doi: 10.1016/j.neuron.2014.10.018

Willner, P. (1986). Validation criteria for animal models of human mental disorders: Learned helplessness as a paradigm case. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *10*(6), 677–690. doi: 10.1016/0278-5846(86)90051-5

Yamashita, Y., & Tani, J. (2012). Spontaneous prediction error generation in schizophrenia. *PLoS ONE*, *7*(5), 1–8. doi: 10.1371/journal.pone.0037843