In press COMMENTARY to Psychological Inquiry

Building a Cybernetic Model of Psychopathology: Beyond the Metaphor

David Rudrauf[a,b,c] & Martin Debbané[a,d]

a. Department of Psychology and Educational Sciences, University of Geneva, Geneva, Switzerland; b. Swiss Center for Affective Science, Campus Biotech, University of Geneva, Geneva, Switzerland; c. Centre Universitaire d'Informatique, University of Geneva, Geneva, Switzerland; dResearch Department of Clinical, Educational and Health Psychology, University College London, London, United Kingdom

Introduction

DeYoung and Krueger's (this issue) proposal of a cybernetic account of contemporary psychopathology strongly resonates with our own views, and we would argue—though this might come as a surprise—with other earlier classical references, which we see as important and somehow neglected. This includes Palo Alto's original systemic approach and Freudian metapsychology, which is itself fundamentally a theory of regulation over objects of desire (Paradiso & Rudrauf, 2012). In accord with DeYoung and Krueger, we hold that psychological science needs generative models of the mind and of its connection to behavior, capable of (a) explaining and predicting the mind and behavior in an operational way, (b) integrating "dimensional approaches" and "a mechanistic account," (c) embedding personality determinants and interindividual variability, and (d)in a manner that can inform "both scientific and clinical thinking." We believe it is essential for the future of precision diagnosis, treatment recommendations, and treatment response monitoring.

Toward this endeavor, we want to emphasize that the word cybernetic, when taken literally, implies theories and models that go beyond "a simple heuristic." Beyond a fertile metaphor, it requires mathematical formalization and computational expression to offer a tangible, operational scientific model, providing interpretable, quantitative parameters that account for expressed behaviors based on underlying psychological mechanisms. The task is demanding and highly technical. It calls for unifying psychology within a general scientific framework (Lagache, 1949), in spite of its splitting into poorly integrated and generally qualitative specialized subfields, on the basis of fully computational principles.

As DeYoung and Krueger explain, a cybernetic model must implement internal "representations" of the world and goals, together with "operators" enacting behavioral strategies. It must be governed by a general algorithm, cycling through mechanisms of (a) goal activation, (b) action selection, (c) action, (d) outcome interpretation, and (e) goal comparison. It thus must integrate a computation of expectations' match and mismatch but also metacognitive processes in order to implement feedback mechanisms, update goals, and adapt them within a global process of optimization.

Important to note, the psychological validity of such principles and the operational value of a cybernetic model require a high level of specification and complexity. To grasp the challenge, it suffices to consider, as stated by the authors, the definition of subjective goals that are remote from the canonical set of evolutionary goals: survival and reproduction, by

the mean of fitness. In humans, these goals more immediately manifest in the context of a struggle for love and recognition and relate to functional and social constructs (Paradiso & Rudrauf, 2012). They may appear at time as incongruent, maladapative or suboptimal, i.e., misfits with respect to direct survival and reproduction or even wellbeing.

Moreover if, as DeYoung and Krueger argue, psychopathology is a "persistent failure to move toward one's goals due to failure to generate effective new goals, interpretations, or strategies when existing ones prove unsuccessful", cybernetic regulatory loops must be "generative," that is, they must entail some kind of learning from experience (Bion, 1962), a healthy dependency to experience for further learning and resilient predictions, as a way to mitigate the risk of psychopathologies.

In this context, part of the struggle for the mind as a cybernetic system, as emphasized by DeYoung and Krueger, is the necessity to model and transitorily absorb entropy, because modeling the complexity and uncertainty of the world—and, we would argue, social relationships—represents a key element to the mind's allostasis. In some general sense, failure to absorb entropy must represent a source of either transient or prolonged stress. The latter acts as a probable cause of anxious and depressive syndromes, because it challenges the mind in its key process of stability through transformations.

However, here the reference to the technical concept of entropy struggles with its generality. If entropy can function as a general "threat" but also sometimes, as acknowledged by the authors with amazement, as a "reward," it cannot in and of itself fully account for how threats or rewards are quantified and integrated in the mind as obstructive or conducive to goals. As we shall see, entropy per se is not in this context the very quantity that needs to be reduced as part of the cybernetic process. Other fundamental quantities from information theory, related to statistical mechanics, need to be considered. Moreover, the actual internal metrics of success for goal pursuit used by individuals, we argue (see below), are highly layered and rely on a variety of affective dimensions, some related to reward and others to punishment. They are attached to complex, often contradictory value systems and call for specific computational operationalizations.

Likewise, to further grasp the scope of the required cybernetic model, appraisal processes have to be conceived as relying on core, active projective mechanisms underlying nonsocial and social perspective taking, in its function of inference and emotion regulation. DeYoung and Krueger are certainly in line with this view since they point out that: « rather than waiting to explore only when entropy increases spontaneously, it is advantageous to explore voluntarily, which means intentionally increasing the entropy of the system, with the expectation that one will be able to reduce it successfully again, having learned new adaptations. »

In this perspective, we argue that to understand how goals can be actively construed as "representations of a desired future state," as part of a global regulatory process, we must offer a theory of subjective experience itself. This does not simplify the task. Subjective experience is an integrative process that frames perception (Rudrauf et al., 2017) and critically conceals multiple perspectives, in relation to past, present, and future states, each laden with affective dimensions and their dynamics. It directly contributes to guide

motivation and potential action/behavioral programs. It heavily relies on the capacity for imagination and social perspective taking as key mechanisms of anticipation and exploration, in which others' perspectives matter. This in our view represents a central component of any valid model of normal and pathological psychology. In this sense, the cybernetic function of the system is effectively articulated to an evolutionary process of sorts, that is of learning from and transmitting information to peers and to the next generation, which builds increasingly complex webs of knowledge contained within the human species (Fonagy & Allison, 2014). According to the model we summarize next, echoing DeYoung and Krueger's cybernetic orientations, the imagination is a projective mechanism, of which the link to emotion and information theory represents the core of the control of nonsocial and social behaviors. The imagination is what renders possible a differentiated process of global (vs. local) optimization of outcomes that can maximize wellbeing and promote resilience in the long run (Rudrauf et al., 2017). Partial to overall failures of this overall process are, we argue, intimately linked to a range of psychopathologies.

But back to our emphasis of the technical nature of the word cybernetic, beyond conceptual claims, how can we formulate subjective experience and its generative relations to other cognitive and affective processes in a formal cybernetic framework?

Let us make some key preliminary epistemological and methodological remarks as they lead us to a paradigm shift in approaching psychological science. We concur with DeYoung and Krueger that an adequate model must operate so as "to begin identifying the underlying causes of dysfunction in each of the major dimensions of psychopathology, in terms of psychological processes that can be functionally unified" and that a model of "psychological dysfunction" is needed. However, perhaps in contradiction with the current mainstream, as well as with some of our past endeavors, we claim that it would be a mistake to seek primary foundations for such scientific model in "complex networks of brain systems," as suggested by the authors . To us contemporary approaches in neuroscience are not the golden road to reveal the secrets of human psychology and build its scientific theory. We claim that a cybernetic and thus computational theory of the normal and pathological mind should be first and foremost built upon purely psychological constructs, and only secondarily (and not necessarily) connected to brain correlates and functions. First, this is because the best of contemporary neuroscience is, at this point, far from offering operational concepts for psychological science: Clinical psychopathology (as opposed to neurology), in its practice, is about interpreting human behaviors based on clinical methodologies and psychological concepts. If brain science may be critical for a complete science of the embodied mind, the current trend in psychology of focusing on neuroscience has, we believe, yielded limited effective results that have revealed truly useful for psychopathology and clinical applications in psychology, in spite of the volume of research. Second, and not the least, the most advanced neuroscience of the mind still encounters methodological and epistemological issues that are deeper than often acknowledged in spite of its apparent scientific rigor (see Marrelec, Messe, Giron, & Rudrauf, 2016; Rudrauf, 2014). References to the brain or to a catalog of brain structures, networks, and molecules cannot be a substitute for psychological and psychopathological model. The level of both details and integration that would be required from brain science to account for psychological processes in an operational way is at this point completely out of reach

(Rudrauf, 2014). We thus believe that such neuroscientific focus can become a true hindrance to the development of a veritable computational psychology and psychopathology. We argue that psychology and psychopathology are better off seeking their scientific foundations in mathematics and computer science (Rudrauf et al., 2017), building their models through the implementation and simulation of artificial agents, based on strong, interpretable artificial intelligence (AI), combined with model selection schemes that can be confronted to empirical data according to the models' predictive power.

Upon these observations and for advancing the discussion on a cybernetic framework for psychology, we summarize herein our own endeavor toward such integrative model.

The Projective Consciousness Model: Integrating Cognition, Emotion, and Imagination in a General Cybernetic Framework

We recently introduced a general mathematical model of the "embodied mind," the projective consciousness model (PCM). It is a basis for developing a psychologically inspired interpretable AI used as a simulable generative model of psychological processes: from perception, imagination, social perspective taking, appraisal, emotion, and motivation to behaviors and their interactions, formulated in a unified cybernetic framework, with well-behaved Bayesian statistics for inference (see Rudrauf et al., 2017, for an introduction to the general principles of the model). The PCM follows the hypothesis that the mind is a process performing active inference (Friston, 2015, 2016) to navigate and learn from its environment in a globally optimal manner, maximizing adaptation and resilience. The PCM builds upon projective geometry and the free energy principle (Friston, 2010). It is compatible with recent Bayesian formulations referring to predictive coding (see Friston et al., 2015, 2017) but does not need to invoke a model of brain implementation. It offers a richer and more comprehensive psychological model than other formal formulations of active inference. It incorporates an explicit model of subjective experience manifesting as a 3dimensional Field of Consciousness (FoC). The FoC frames perceptual and imaginary inference in the context of multivariate appraisal and motivational mechanisms for the selection and orientation of action, in a manner that can be simulated in variety of virtual test environments.

The FoC is structured, as a spatial structure, by a three-dimensional projective space (a space in perspective) relating a point of view to a representation of the world, including of the self and others. The FoC undergoes projective transformations, that is, geometrical transformations relating actual and possible points of view in perception and imagination. Projective transformations implement key spatial components of non-social and social perspective taking. The FoC also operates as a force field, which drives intentionality and attention for orientation and action selection, under the control of a process of global optimization based on free energy (FE) minimization. FE is an upper-bound on surprise in an Information theoretic sense and quantifies the dissatisfaction of expectations in relation to prior beliefs and preferences, incorporating the divergence of current priors from sensory evidence and the negative entropy of the information to integrate.

FE minimization and appraisals. In the PCM, FE is quantified over fields of conditional probabilities, which are encoded in memory and normally updated as a result of action and

ensuing sensory evidence. These probabilities notably express expected appraisals of values as a function of possible actions, past experience, and more generally time and space. For instance, assuming that hedonic reward and safety are part of the prior preferences of the system, a high probability of hedonic reward $\mathrm{P}(\mathrm{hedonic}\,|\mathrm{action},\mathrm{T}) = 1$ or safety $\mathrm{P}(\mathrm{safety}\,|\mathrm{action},\mathrm{T}) = 1$ following an action under a given perspective T implies a low FE and weighs in favor of that action. On the contrary, a low probability of hedonic reward $\mathrm{P}(\mathrm{hedonic}\,|\mathrm{action},\mathrm{T})$ or safety $\mathrm{P}(\mathrm{safety}\,|\mathrm{action},\mathrm{T})$ following an action under a given perspective T implies high FE and weighs against that action. An important point underlying the richness and complexity of the generative mechanism is that these conditional probabilities represent multiple appraisal dimensions (e.g., hedonic, safe, or norm compatible), which are evaluated simultaneously by the FoC, in spite of their possible contradictory values, and which the system tries to globally optimize, to maximize the likelihood of desired outcomes.

FE minimization and perspective taking. FE minimization drives the choice of projective parameters used for perspective taking, locally in perception and remotely in imagination (e.g., first-person vs. third-person perspective taking). It further defines the selection of actions (which can include simple displacements, complex actions, or the absence of an action), under a variety of perspectives (standpoints, directions of aiming, spatial scope), affording different focalizations and weights on the distribution of information. Combined with priors, perspectives can be attributed to self or others, related to factual or counterfactual information, and to a combination of representations of the past, present, and future, altogether driving the intentional and affective states of the AI.

The distribution of FE as framed by the FoC is integrated in a statistical manner across space and time to explore and exploit its spatial and temporal gradients. At each instant, the PCM computes several FoC (perceptual and imaginary), relating past, present, and future situations (like short subjective sequences). It computes an overall FE summary statistic across these series of FoC, which functions by analogy as a general gut feeling about their information content.

The **global optimization** of the FoC sequences across time drives the PCM behavior. Globally optimal perspectives $\mathrm{T_i}$ on the distribution of information and decisions of $\mathrm{action_i}$ are selected so that:

$$(\mathrm{T_i}, \mathrm{action_i}) = \mathrm{argmin}(\mathrm{FE}(\mathrm{preference}\,|\mathrm{T},\mathrm{action})) \quad [1]$$

Computationally, perspective taking allows the agents to escape local minima of FE (Rudrauf et al., 2017). PCM agents thus attempt to optimize the satisfaction of their preferences by considering global solution spaces (e.g., through projective imagination), which offer a bigger potential for resilience (through the consideration of alternative paths of action) than purely local solutions (e.g., through immediate perception). Daily experience functions as a constant challenge to the cybernetic system, which is continuously obliged to transform and optimize goals in order to remain stable.

Motivation and action programing. Estimates of optimal projective transformations and actions $(\mathrm{T_i}, \mathrm{action_i})$ are then used for action programing. At the behavioral level, the axis of

approach versus avoidance behaviors can be modeled straightforwardly. Generally speaking, if the anticipated change of FE, $\Delta\mathrm{FE}$, as compared to the current or a past situation, is inferior to zero, then such an amelioration weighs in favor of approach behaviors. If it is superior to zero, on the contrary, then the anticipated aggravation weighs in favor of avoidance behaviors. Important to note, $|\Delta\mathrm{FE}|$ yields a general motivational force proportional to the anticipated amelioration or aggravation of the satisfaction of prior preferences and can be seen as a proxy for resource mobilization in relation to the programing of actions. Thus, desired states of the agent correspond to states that minimize FE, and the drives toward these states (respectively, the pull away from them) are proportional to the actual or anticipated decrease (respectively, increase) of FE. Goals are explicit representations of locations, action outcomes, and states of the spatialtemporal world, associated with values that minimize FE.

Outcome variables. The overall algorithm in its current development can output a variety of parameter estimates:

(a) internal analytics of affective states (appraisals, emotions), (b) motivational and physiological parameters (Autonomic Nervous System estimates), and (c) behaviors (overt attention, trajectories, approach-avoidance, facial expressions).

Perspective taking, appraisal, and resilience to stressors: Simulations. In Figure 1.1, we show PCM simulations of the role of imaginary perspective taking toward the achievement of future goals in the motivation of facing a stressful challenge of crossing a pit. Likewise, simulations of social perspective taking can be based on a direct extension of the core algorithm, whereby agents take perspective from possible points of view of others, on the environment or on themselves, and attribute prior beliefs to them.

From the General Algorithm to Specific Emotional States and Their Regulation

The PCM algorithmic architecture explicitly formalizes the relationships between appraisals, the elicitation of specific emotions, their expression, and mechanisms of regulation. Building upon recent proposals (Cunningham, Dunfield, & Stillman, 2013; Joffily & Coricelli, 2013), we can begin to show how the internal dynamics of FE in the PCM can be analyzed to develop a quantitative understanding of affective processing and emotion expression as part of the larger process of active inference. Joffily and Coricelli (2013) offered a first formal model of interpretation of general FE dynamics as relating to the dimension of valence with respect to factual and epistemic evaluations but in a manner that was limited in scope, and with a narrower understanding of FE, and in a context that was difficult to generalize. Notably, they did not consider the presence of multiple dimensions of appraisal and their relation to FE and did not relate FE dynamics to a general psychological model of intentionality, or to a model of motivation and action, necessary to further develop emotion theory in relation to complex goal-directed behaviors.

To understand emotional states, we must consider complex intentional stances relating (a) past and present, actually experienced FoCs; (b) past and future anticipated (recalled or imagined) FoCs; (c) across multiple layers of appraisal simultaneously (e.g., hedonic, safe and norm compatible). The dynamical profiles of the terms of FE associated with a set of

such FoCs, then offer a basis to understand the clustering of emotional states into constructs such as basic emotions and more sophisticated models of affective categories (Figure 1.2).

For instance, when considering an appraisal dimension of safety alone (Figure 1.2a, left tier), a growing FE from past to (anticipated) future can be interpreted as a sense of growing anxiety based on the recall of safer places, a current predicament, and an anticipated aggravation of circumstances. When considering an appraisal dimension of hedonic pleasure alone (Figure 1.2a, right tier), a growing FE from past to (anticipated) future can be interpreted as a depressive mood, based on the recall of better times, a current lack of pleasure, and an anticipated state of anhedonia associated with a high level of reward-related FE (which implies a low state of reward).

We note here how essential it is to consider the problem of FE minimization for a PCM agent as a multivariate problem, in which a multiplicity of motives and perspectives, possibly quite contradictory, operate simultaneously and with complex conditional relationships. FE is computed over all n layers of appraisal incorporated in the agent simultaneously, and the behavior of the agent is driven by the spatial and temporal gradients of the weighted sum of all such specific contributions FE_i to FE:

$$\mathrm{FE} = \sum_{i=1}^{n} \omega_i FE_i \qquad [2]$$

The multiplicative weight ω_i sets the current weight granted to a given appraisal dimension in the overall dynamics of FE integration. For instance, combining two dimensions of appraisal such as safety and hedonic reward (Figure 1.2b, left tier), and assuming an equal ω_i between them ($\omega_{\mathrm{safety}} = \omega_{\mathrm{hedonic}}$), a combined decreasing FE from past to future can be interpreted as an epistemic hope, starting from a suboptimal state and anticipating a global amelioration of circumstances across both dimensions. If these two dimensions are moving in opposite directions, for example, a decreasing FE related to hedonic reward and an increasing FE related to safety (corresponding to a lack of safety), the agent is locked into a state of ambivalence or mixed feeling, with FE terms cancelling each other and offering no path of minimization, thus no rule for action, neither in terms of approach nor in terms of avoidance.

Adaptive and maladaptive emotion regulation can be modeled in this context quite straightforwardly by modulating ω_i. A strategy of emotion regulation to resolve the FE deadlock could be, for instance, to tune the relative weight of the dimension of safety versus hedonic reward on the overall appraisal. Because a PCM agent attempts to minimize FE this would motivate the agent to act accordingly. A strategy leading to $\omega_{safety} \gg \omega_{hedonic}$ would lead to a high sensitivity to anticipated danger, with an overall FE increasing, and to potentially maladaptive avoidance behaviors, as, for instance, in non-social or social anxiety. On the contrary, a strategy leading to $\omega_{safety} \ll \omega_{hedonic}$would make a growing sense of danger be overridden by the growing anticipated pleasure, and overall FE would decrease as a result, leading to potentially maladaptive approach behaviors, such as in risk-taking behaviors.

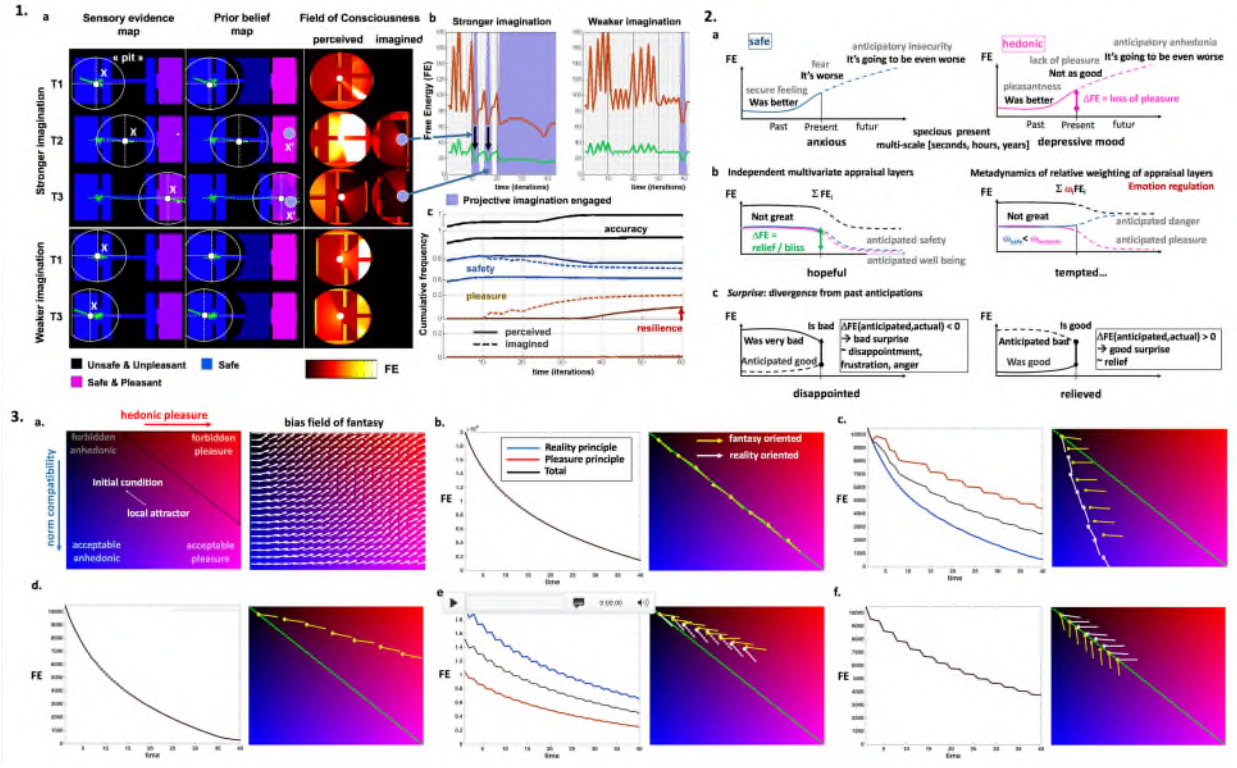

Figure 1. Figure 1. 1. Simplified PCM simulation of a stressor challenge. a. Left Tier. Maps of the state of the world model (as a 2D finite plane) and agent. Rectangular maps (left). 2D three-room-environment, with a central challenge room simulating an unsafe and unpleasant virtual pit (black), a safe but otherwise neutral departure room (left), and a pleasant and safe goal room (right). Sensory evidence maps represents the factual expectation of valenced events (see color code). Prior beliefs maps, the subjective beliefs of the agent. Large white circles represent the scope of the Field of Consciousness (FoC) on the maps. Two types of agents are presented: one with strong imagination drives, more likely to imagine remote solutions, one with weaker imagination drives. Rows correspond to successive time periods. Right Tier. Circular maps. FoC of the agent (current perceived or imagined contents of consciousness). The color represents free energy levels (the darker the lower). When the agents engage in remote projective imagination both the local and remote layers of the FoC are shown. b. Time course of FE for the two agents. Optimal perspective taking across possible first person perspectives (i.e., projective transformations centered at the agent location) has the lowest average FE (green line) versus grand average (red line). c. Cumulative frequencies of perceived and imagined experiences undergone by the agent across time, with appraisal of safety and pleasure connecting FE to affective dynamics. Accuracy of beliefs is also represented. The agent with a stronger imagination drive faces the challenge after hesitating to cross the pit and reaches the goal room, which maximizes utility. It uses goal-related projective imagination, in an optimistic way, which reduces anticipated FE, to overcome the challenge, and arrives in the goal room. Its gain in perceived pleasure can be used as a measure of resilience to initial conditions (red arrow), which were safe but anhedonic in nature. Through exploration and prior updating, its accuracy increases. The agent with a lower imagination-drive never faces the challenge, does not explore its environment, and never enjoys a better condition and greater accuracy, but wanders in a repetitive manner in its local environment. 2. Emotional states from FE dynamics (see text in corresponding section). 3. PCM cybernetics and Psychodynamics. Simulations of simple PCM agents driven by hedonic pleasure (HP) and norm compatibility (NC) appraisals. a. Left: Flatland world model. Color code indicates combinations of probabilities of HP and NC. In these simulations, maps related to sensory evidence and prior beliefs and preferences are identical. Only the relative weight of HP and NC, and that of the driving force exerted by fantasy and reality on action and imagination are manipulated. Agents are represented with vectors indicating their current location and intentional directions. Right: Example of bias field for fantasies driven by HP. b. “Normopath.” Left: FE as a function of time (simulation iterations). Agents always try to minimize FE, along all dimensions (FE related to norm compatibility in blue; FE related to hedonic pleasure in red; overall FE in black). Right: Fantasy (yellow) and reality (white) oriented vectors indicating imaginary and real travels. In all cases presented here, fantasy and reality influence each other but can be dissociated in the model. The agent walks the line (green) of balance between pleasure and norm compatibility, seeking both pleasure and norm satisfaction in fantasy and reality, whereas both norms and pleasure exert a driving force on real action and imagination. (Parameters: w(pleasure|reality) = 1; w(norms|reality) = 1; w(pleasure|fantasy) = 1; w(norms|fantasy) = 1 ; w(reality | perception) = 0.5; w(reality | imagination) = 0.5; w(fantasy | perception) = 0.5; w(fantasy | imagination) = 0.5). c. The agent seeks norm compatibility in perception and pleasure in imagination, reality drives actions against fantasy. The agent is driven towards norm compatible but anhedonic contexts in perception and action (white vectors), but has fantasies oriented towards hedonic contexts (yellow vectors). FE related to the principle of reality is minimal (satisfaction) and FE related to the principle of pleasure is high (unsatisfaction) (Parameters:

w(pleasure|reality) = 1; w(norms|reality) = 1; w(pleasure|fantasy) = 1; w(norms|fantasy) = 0.1 ; w(reality | perception) = 0.9; w(reality | imagination) = 0.1; w(fantasy | perception) = 0.9; w(fantasy | imagination) = 0.1). d. The agent seeks pleasure and care little for norm compatibility. (Parameters: w(pleasure|reality) = 1; w(norms|reality) = 0.1; w(pleasure|fantasy) = 1; w(norms|fantasy) = 0.1 ; w(reality | perception) = 0.5; w(reality | imagination) = 0.5; w(fantasy | perception) = 0.5; w(fantasy | imagination) = 0.5). e. The agent seeks equal pleasure and norm compatibility in reality but fantasy is dominated by pleasure, and norms drive actions with less potency than fantasy. The agent drifts towards forbidden pleasures both in imagination and action, but orients its perception toward norm compatible regions, thus featuring an active dissociation between assumed perception and imaginary preferences, not unlike the constitution through denial or repression of an unconscious. It is hindered in its progression by these contradictory motives (Parameters: w(pleasure|reality) = 1; w(norms|reality) = 1; w(pleasure|fantasy) = 1; w(norms|fantasy) = 0.1 ; w(reality | perception) = 0.9; w(reality | imagination) = 0.1; w(fantasy | perception) = 0.9; w(fantasy | imagination) = 0.1). f. This time the agent seeks norm compatibility in imagination, and pleasure in reality, while it is equally driven by imagination and perception. The agent is hindered in its progression by contradictory motives. It walks the line, but its attention in perception and in imagination is divided. (Parameters: w(pleasure|reality) = 1; w(norms|reality) = 0.1; w(pleasure|fantasy) = 0.1; w(norms|fantasy) = 1 ; w(reality | perception) = 0.5; w(reality | imagination) = 0.5; w(fantasy | perception) = 0.5; w(fantasy | imagination) = 0.5).

Beyond departure from preferences across layers of appraisal, violations of expectations between prior-based, anticipated (imagined) states and sensory-evidence-based experienced (perceived) states constitute a central contribution to the dynamics of FE and to the process of revision of priors (Figure 1.2c). The difference $|\Delta \mathrm{FE}|$ in FE between imagined and perceived states after action quantifies a general level of surprise (which is closer to the standard narrower understanding of FE). When one considers the direction of changes of $|\Delta \mathrm{FE}|$ (increasing or decreasing), one obtains valenced states that can be identified with "bad surprise," when $|\Delta \mathrm{FE}| > 0$ (e.g., as underlying disappointment, frustration or anger), or "good surprise," when $|\Delta \mathrm{FE}| < 0$ (as underlying relief).

This multivariate quantification of FE across dimensions of appraisal and epistemic inference (perception vs. imagination) subsumes concepts of valence and arousal (which can be directly related to FE), categories of affective states (fear, anxiety, sadness, satisfaction, joy, hope, surprise, disappointment, relief), the embedding of simple emotional states into more complex ones (e.g., hope, anxiety, depressive moods), which directly emerge from the intentional model at play in the PCM and its relation to the minimization of FE. The differentiation of such intentional models and affective states is central to normal and pathological development.

Important to note, in the PCM, the affective states are also directly related to motivational parameters that can be expressed as physiological dimensions and control-signals for effectors (e.g., musculoskeletal systems). The outcomes of these processes yield new sensory evidence, entering the cybernetic process of active inference. In other words, the state of the body of the agent in its relations to the world and to others is constantly quantified by the process of active inference and becomes a space of appraisal from multiple perspectives, combining imagination and perception. This can be interpreted as an embodied projective self-model, which itself becomes a prior, with a variety of possible levels of flexibility or rigidities (Rudrauf et al., 2017).

Thus, the model sheds light, in a computationally tractable way, on the workings of emotion in the larger ecology of embodied thought and cognition. It offers a framework to study and quantify the interplay between (a) prior beliefs, rigid or labile, undifferentiated or differentiated, across a multiplicity of appraisal dimensions; (b) the ability to deploy projective imagination and take a multiplicity of alternate perspectives (possibly attributed

to others) on these beliefs (note that the ability to distinguish one's own priors from those attributed to others is essential (Fotopoulou & Tsakiris, 2017) in the process and can become an object of quantitative inquiry with the PCM); and (c) the capacity to update priors based on new sensory evidence or to tune the relative weight of those priors (e.g., preferences) adaptively or maladaptively, in the appraisal process.

Within this manifold, one can start modeling the broadest range of emotional and conative experiences, from (a) "psychic equivalence" (Fonagy & Target, 1996), in which intentional states can appear as an unquestionable judgment of truth about the world, self, and others, for instance, with rigid prior beliefs and an egocentric projective process (as in paranoia), to (b) "mentalized affectivity" (Fonagy, Gergely, Jurist, & Target, 2002), by which key learning points (reformulation of goals) work through experiences of emotions that can be experienced as emotionally meaningful, and thought about from multiple perspectives. Complex internal affective clusters and their impact on behavior can be understood, including states of ambivalence, which can be the condition for the development of adaptive and maladaptive strategies.

Perturbations of the Functional Architecture and Generative Models of Psychopathology

The internal world model built through active inference by an optimal PCM maximizes the satisfaction of the agent's preferences, tends to secure its resilience to adverse events, and most important sustains an active learning mechanism that tends to secure stability through change (i.e., allostasis). Perturbations of the functional architecture, at the level of both prior encodings and core computational mechanisms (from learning rate to capacities for perspective taking), can be integrated in a graded manner to generate dysfunctional processes and behaviors, yielding a parametric range of pathological agents with outcome variables that can be compared. We can consider, for example, general perturbations of the degree to which the optimization algorithm relies on local versus global optimization. For instance, we can limit global optimization by impairing projective imagination, yielding agents stuck in local optimization generating a variety of compulsive and anxious-like disorders. On the contrary, agents pushing global optimization beyond the capacity of their executive functions can set goals that are impossible for them to reach, yielding a permanent, potentially pathological state of dissatisfaction, which can generate states similar to exhaustion and despair. A variety of adaptive and maladaptive coping mechanisms can be manipulated. Forms of denial can be generated by weighting down sensory evidence integration in a manner that makes agents avoid confrontation with reality. Rigid, paranoid prior beliefs can be parameterized by playing on priors initial encoding and down weighting specific prior updating rules.

Perturbations of specific appraisals, combined with perspective-taking styles, either through up- or down- regulation, can make agents generate behaviors related to a range of non-social and social anxiety states (e.g., through the overweighting of safety concerns with negative expectations making safety related FE goes up), or depressive anhedonic states (e.g., through the up-weighting of hedonic reward concerns with negative expectations). Depression, for instance, has been related to a psychopathological cascading maladaptive reaction to negative life events or perceptions of loss, which then can lead to the expression of symptoms across a number of domains including affective, cognitive, and somatic

symptoms, as well as disruptions in social processes and relationships. In the PCM, an overweighting of an appraisal of potential loss of objects, which are believed by the agent to be essential to its FE minimization, is expected to foster avoidance behaviors toward situations otherwise sources of positive rewards. This can thus mechanically hinder exposure to key learning experiences and reappraisal mechanisms and induce a vicious circle, leading to states similar to helplessness. Trauma can be thought of as the result of an extreme negative surprise leading to an extreme updating of negative priors. The encoding of a strong negative prior then imposes its weight on future appraisals of situations and can mechanically lead to recurrent reminiscences and biased interpretations due to, for instance, an over-interpretation of sensory evidence under the influence of the negative prior.

More generally, personality traits and tendencies can be modeled as part of the set of priors that the individual agents embed, which include the agents' epistemic trust toward the world and others. Agents can be highly personalized, in connections to a wealth of empirical data from multiple sources, for instance, by analyzing operational connections between the PCM architecture, mechanisms and priors, and dimensional frameworks such as the Research Domain Criteria initiative (see Krueger & DeYoung, 2016). Thus the PCM framework is congruent with DeYoung and Krueger's (this issue) aim of combining cybernetics and dimensional approaches. All of these exemplars are currently being modeled and developed within the PCM framework.

Psychodynamic Implications of the PCM

The model of subjectivity embedded in the PCM is nontrivial. As an illustration of this notion, it is highly relevant to emphasize how the PCM implies mechanisms that directly resonate with Freud's first and second topic models (Freud, 2013, 2015). In the PCM, the imagination is analogous to a function of fantasy and (day) dreaming, and multivariate FE minimization across layers of appraisal such as hedonic pleasure, safety, and norm compatibility embeds a Principle of Pleasure and a Principle of Reality in connection to the Bayesian interplay of prior beliefs and sensory evidence. The process implements a cybernetic mechanism of minimization of tension, prone to contradictory motives and necessary compromises. The agent is literally a divided subject. The optimization of the process can make regions of an agent's memory, and entire sets of possible representations of goals act at the same time: (a) repulsively vis-a-vis conscious access and enaction, for instance, due to layers of safety and norm compatibility, and (b) attractively, for instance, along layers of hedonic (though perhaps forbidden) pleasures. Freud postulated that in fantasy (imaginary representations) and dreams, a lower weight was generally placed on norm compatibility and safety, that is, on censorship, allowing the mind to derive virtual satisfaction from forbidden pleasures.

As shown in Figure 1.3, simple models of PCM agents may thus tend to avoid confrontation with these regions of forbidden pleasure in their memory, for instance, through perception and action or by restricting imagination (in particular when safety and norm compatibility have a regulatory weight that balances hedonic pleasure). Such process is analogous to repression and can manifest behaviors similar to compulsive repetition, with an agent attracted by but trying to avoid these regions. Agents with various relative weights of

pleasure and norm compatibility in their appraisal, and various relative weights of reality check (sensory evidence) versus fantasy (projective imagination) in the motivation of their actions, will display complex and sometimes paradoxical or dissociative behaviors. This reflects the compromise made by such active inference systems in satisfying all the motives that drive them at once.

Even under this simplistic implementation of the PCM, which admittedly is in its infancy in terms of development, complex states and behaviors emerge from the interplay of motives, the dissociation between perception and imagination, as well as biases on action and orientations. The choice of parameters, the tuning of beliefs and projective mechanisms, can implement and allow investigators to study how personality traits interact with mental states and behaviors. Likewise, the PCM framework opens the possibility to investigate psychodynamic principles and their generative role in the emergence of pathological behaviors (CarhartHarris & Friston, 2010), as well as to test their validity based on quantitative predictions, and paves the way toward a computational metapsychology.

Concluding Remarks: Psychopathology and Cybernetics

We have underlined a number of similarities we find between DeYoung and Krueger's (this issue) call for a cybernetic model of psychopathology and our own views on psychopathology and modeling, which are formulated around the PCM and informed by recent developments in the mentalization-based framework (Fonagy & Allison, 2014). We have stressed three points of possible differences.

The first point is that the cybernetic proposal presented by DeYoung and Krueger is not as such computational, and propositions of operators appear to promote neuroscientific rather than psychological models of the mind. In our view, critically, operationalizations must rely on fine phenomenological description that account for the intrapsychic and intersubjective levels, and one of the key psychological mechanisms that permits this is the imagination. We have expressed why we feel that psychopathology models need to be resolutely psychological in nature, and why we are required to provide psychological models that can be formalized by mathematics and implemented as simulations of artificial agents. We then summarized the elements of the projective consciousness model that starts addressing the need for mathematical formalization, and the integration of the imagination as a key psychological process for sustaining mental health.

Our second point is the hazards that a cybernetic model of psychopathology may encounter if only framed as self-regulating system. We insisted on the role of social perspective taking as part of the core function of the imagination, for perceptual inference and global optimization of the cybernetic process. The basic condition of infant growth poses, in our view, the condition that any model must be compatible with a developmental view of human self-regulation, which makes it dependent upon regulations that come from external agents. Human infants appear to be preprogramed to engage in early communicative interactions with attachment figures in order to survive but also to derive their first representations about the environment (Csibra & Gergely, 2011) and decipher trustworthy and untrustworthy sources of knowledge (Corriveau et al., 2009; Fonagy & Allison, 2014). Self-regulation, mental health, or, generally speaking, adaptive fitness and social functioning

are possible only if developed within sufficiently stable caregiving relationships (Groh et al., 2014).

Taking into account the nature of the development of a self-regulating system within its social context, we can revisit the idea that an operationalized cybernetic system can be at risk for psychopathology when failing to generate new goals. The reasons for this failing can be numerous, but where they all seem to converge is that they ensue an incapacity for the subject to "take in" from the social environment and its resources, to learn from other sources of knowledge, which we translate as failures to learn from experience. In operational terms, the system effectively fails to generate new goals, due to a roadblock between self and the world. What an operationalized cybernetic approach affords is the opportunity to analyze the many ways the subject fails to take advantage of the environment and how that translates into a symptomatic failure to generate new goals. Therefore, the failure to generate new goals can be said to be the symptom of a system that, for a number of possibly different reasons, ranging from thwarted neurodevelopment to trauma, cannot successfully and confidently engage with sources of knowledge to pursue learning. In most psychopathology, the individual withdraws from the influence the environment can have on transforming her or his priors into new experiences, and eventually new priors.

In our view, the generation of new goals is intimately tied to transactions with the environment, in particular the social environment, because it is what is understood from the environment that modulates free energy. The process heavily relies on projective mechanisms at the core of subjectivity, allowing the system to take a variety of perspectives for appraisal and reappraisal of priors. This is why we feel that subjectivity must lie at the heart of the model, as it conditions the relationships between novel experience, active inference, goal activation, monitoring processes, and generation of new goals.

This brings us to our third point of potential difference, which concerns the nature of psychopathology. DeYoung and Krueger differentiate between models of psychopathology that are rooted in evolutionary function and their model of psychopathology, which would be rooted in "persistent failure to move toward's one's goals, due to failure to generate effective new goals" . Although we agree that simplistic evolutionary proposals miss essential psychological categories, we would argue that a cybernectic model of psychopathology needs to be expressed within an evolutionary framework, which would be defined not only in terms of direct survival, reproduction, and fitness needs but rather in terms of the almost unique human evolutionary dynamic to transfer knowledge from one generation to the next, which itself acts as a mechanism of selection, over and above the direct transfer of genes. We thus claim that if the human brain is designed to generate a mind that performs active inference in a projective manner, its function within the context of its species is to develop evermore complex systems of understanding that contain information that can be passed on to others and create broader mechanisms of resilience, including trans-generationally.

In concluding, we wish to emphasize that, owing to its computational nature, the PCM can be used to differentially predict behaviors based on underlying mechanisms that are interpretable from a clinical and psychopathological standpoint. These mechanisms can be

used, in turn, to derive new hypotheses about the potential failure of the cybernetic system to generate new goals and inform possible treatment interventions. Combined with empirical data, the PCM can become a scientific instrument, integrating interpretable artificial intelligence with clinical observation and procedures. It could then offer a basis, combined with model selection and reverse inference schemes, for the development of meaningful tools aimed at assisting clinicians in precision diagnosis, treatment recommendations, and treatment response monitoring. In other words, we believe that models such as the PCM are the type of methods that can effectively address DeYoung and Krueger's call for a cybernetic framework for psychology and psychopathology.

References

Bion, W. R. (1962). Learning from experience. London: William Heinemann. [Reprinted London: Karnac Books]

Carhart-Harris, R. L., & Friston, K. J. (2010). The default-mode, ego-functions and free-energy: A neurobiological account of Freudian ideas. Brain: A Journal of Neurology, 133(Pt 4), 1265–1283. doi: 10.1093/brain/awq010

Corriveau, K. H., Harris, P. L., Meins, E., Fernyhough, C., Arnott, B., Elliott, L., … de Rosnay, M. (2009). Young children's trust in their mother's claims: Longitudinal links with attachment security in infancy. Child Development, 80(3), 750–761. doi:10.1111/j.14678624.2009.01295.x

Csibra, G., & Gergely, G. (2011). Natural pedagogy as evolutionary adaptation. Philosophical Transactions of the Royal Society B: Biological Sciences, 366(1567), 1149–1157. doi:10.1098/rstb. 2010.0319

Cunningham, W. A., Dunfield, K. A., & Stillman, P. E. (2013). Emotional states from affective dynamics. Emotion Review, 5(4), 344–355. doi:10.1177/1754073913489749

Fonagy, P., & Allison, E. (2014). The role of mentalizing and epistemic trust in the therapeutic relationship. Psychotherapy, 51(3), 372–380. doi:10.1037/a0036505

Fonagy, P., & Target, M. (1996). Playing with reality: I. Theory of mind and the normal development of psychic reality. The International Journal of Psychoanalysis, 77(Pt 2), 217–233.

Fonagy, P., Gergely, G., Jurist, E. L., & Target, M. (2002). Affect Regulation, Mentalization, and the Development of the Self. New York: Other Press.

Fotopoulou, A., & Tsakiris, M. (2017). Mentalizing homeostasis: The social origins of interoceptive inference. Neuropsychoanalysis, 19(1), 3–26. doi:10.1080/15294145.2017.1294031

Freud, S. (2013). The interpretation of dreams (vols. 4 and 5). London: Hogarth Press.

Freud, S. (2015). Beyond the pleasure principle. Psychoanalysis and History, 17(2), 151–204. doi:10.3366/pah.2015.0169

Friston, K. (2010). The free-energy principle: A unified brain theory? Nature Reviews. Neuroscience, 11(2), 127 doi:10.1038/nrn2787

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. Neural Computation, 29(1), 1–49. doi:10.1162/NECO_a_00912

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. Cognitive Neuroscience, 6(4), 187–214. doi:10.1080/17588928.2015.1020053

Groh, A. M., Fearon, R. P., Bakermans-Kranenburg, M. J., Van IJzendoorn, M. H., Steele, R. D., & Roisman, G. I. (2014). The significance of attachment security for children's social competence with peers: A metaanalytic study. Attachment & Human Development, 16(2), 103–136. doi:10.1080/14616734.2014.883636

Joffily, M., & Coricelli, G. (2013). Emotional valence and the freeenergy principle. PLoS Computational Biology, 9(6), e1003094.s.

Krueger, R. F., & DeYoung, C. G. (2016). The RDoC initiative and the structure of psychopathology. Psychophysiology, 53(3), 351–354. doi: 10.1111/psyp.12551

Lagache, D. (1949). L'unite de la psychologie; psychologie experimentale et psychologie clinique.

Marrelec, G., Messe, A., Giron, A., & Rudrauf, D. (2016). Functional connectivity's degenerate view of brain computation. PLoS Comput. Biol, 12(10), e1005031 doi:10.1371/journal.pcbi.1005031

Paradiso, S., & Rudrauf, D. (2012). Struggle for life, struggle for love and recognition: The neglected self in social cognitive neuroscience. Dialogues in Clinical Neuroscience, 14(1), 65–75.

Rudrauf, D. (2014). Structurefunction relationships behind the phenomenon of cognitive resilience in neurology: Insights for neuroscience and medicine. Advances in Neuroscience, 2014, 1. doi:10.1155/ 2014/462765

Rudrauf, D., Bennequin, D., Granic, I., Landini, G., Friston, K., & Williford, K. (2017). A mathematical model of embodied consciousness. Journal of Theoretical Biology, 428, 106–131. doi:10.1016/ j.jtbi.2017.05.032