Evolution, Culture and Computation in Psychiatry

Axel Constant Pruvost
A thesis submitted to fulfil requirements for the degree of Doctor of Philosophy
Faculty of Arts and Social Sciences
The University of Sydney
2023

THE UNIVERSITY OF SYDNEY

This is to certify that to the best of my knowledge, the content of this thesis is my own work.

This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

_____

Axel Constant Pruvost

May 13th, 2023

Published chapters

Chapter	Chapter published as	Contribution	Copyrights
Chapter 1	**Constant, A.** (2021). The free-energy principle: it's not about what it takes, it's about what took you there. Biology & Philosophy, 36(2), 10. https://doi.org/10.1007/s10539-021-09787-1.	I wrote the draft and made the revisions.	Permission granted for thesis
Chapter 2	**Constant, A.,** Clark, A., Kirchhoff, M., & Friston, K. J. (2020). Extended active inference: Constructing predictive cognition beyond skulls. *Mind & Language*, *mila.12330*. https://doi.org/10.1111/mila.12330	I was the principal contributor of the draft and of the revisions.	Open Access
Chapter 3	**Constant, A.,** Badcock, P., Friston, K., & Kirmayer, L. J. (2022). Integrating Evolutionary, Cultural, and Computational Psychiatry: A Multilevel Systemic Approach. *Frontiers in Psychiatry / Frontiers Research Foundation*, *13*, 763380. https://doi.org/10.3389/fpsyt.2022.763380	I was the principal contributor of the draft and of the revisions.	Open Access
Chapter 4	**Constant, A.,** Hesp, C., Davey, C. G., Friston, K. J., & Badcock, P. B. (2021). Why Depressed Mood is Adaptive: A Numerical Proof of Principle for an Evolutionary Systems Theory of Depression. *Computational Psychiatry (Cambridge, Mass.)*, *5*(1), 60–80. https://doi.org/10.5334/cpsy.70	I was the principal contributor of the draft and of the revisions.	Open Access

Acknowledgement

I would like to thank my mother, my friends, and my partner as well as my colleagues and mentors including Paul Griffiths, Paul Badcock, Karl Friston and Laurence Kirmayer for their support; support without which this work would not have been possible.

Abstract

This thesis develops an approach to integrate evolutionary, cultural, and computational approaches to psychiatry in 4 chapters. The claim at the core of this thesis is that a principled holistic explanation of mental disorders would benefit from the integration of explanations in computational, cultural, and evolutionary psychiatry. The argument is presented through two models. The first model is presented in chapter 3, and functions as an ontology of mental disorders that integrates principles of evolutionary, cultural, and computational psychiatry. The second model is presented in chapter 4 and implements this integrative view with a computational model of major depressive disorder. The models that I propose are based on two important philosophical assumptions about *active inference*, the formal theory that underwrites them. First, the two models assume that active inference — and implicitly the free-energy principle — can be applied to the behaviour of non-living systems. Second, the models assume that the cognition and behaviour (e.g., action, perception, and learning) of living systems — such as modelled under active inference — have a formal equivalent in non-living systems. This allows us to apply the free-energy principle to the dynamics of systems that involve nonliving components such as enculturated humans embedded in a material environment. The first portion of this thesis contained in chapters 1 and 2 defends these two assumptions. The second portion of this thesis contained in chapter 3 and 4 presents the two models.

Table of contents

Thesis introduction

"Here one may certainly admire man as a mighty genius of construction, who succeeds in piling an infinitely complicated dome of concepts upon an unstable foundation, and, as it were, on running water. Of course, in order to be supported by such a foundation, his construction must be like one constructed of spiders' webs: delicate enough to be carried along by the waves, strong enough not to be blown apart by every wind. As a genius of construction man raises himself far above the bee in the following way: whereas the bee builds with wax that he gathers from nature, man builds with the far more delicate conceptual material which he first has to manufacture from himself. In this he is greatly to be admired, but not on account of his drive for truth or for pure knowledge of things. When someone hides something behind a bush and looks for it again in the same place and finds it there as well, there is not much to praise in such seeking and finding. Yet this is how matters stand regarding seeking and finding "truth" within the realm of reason."

-- Truth and lies in an extra moral sense, Nietzsche

This dissertation is an attempt at integrating three distinct ways of thinking about mental disorders in psychiatry: (i) as developmentally aggravated vulnerabilities understood as proximate causes shaped by ultimate, evolutionary causes; (ii) as behavioural patterns causing psychological distress and functional impairment configured at the subjective level and shaped by socio-normative causes; and (iii) as suboptimal inference of perception and action caused by lesioned or atypically learned model parameters. These three ways of thinking about mental disorders belong to three different approaches to psychiatry: the evolutionary approach, the cultural approach, and the computational approach.

The central claim of this dissertation is that: *A principled holistic explanation of mental disorders (i.e., an explanation based on a reasoning pattern that accounts for the dynamics of the mind understood as a system made of functionally related parts) would benefit from the integration of explanations in computational, cultural, and evolutionary psychiatry.* I will defend that claim in 4 chapters, which correspond to four published and peer-reviewed articles.

Within the context of this dissertation, however, these articles should not be seen as independent, but rather, as steps on my way to arguing the main claim of this dissertation. How this will be done will be clarified, both in this introduction, and at the beginning and end of each chapter, in the form of introductions and conclusions that join the narrative. I will use interim conclusions to pursue some of the discussions I could not pursue in the articles, or to respond to some of the recurrent problems my colleagues and mentors pressed me on, with respect to the overall project.

Note that I wrote the four chapters of this dissertation in an order different from that of their presentation in this dissertation. I first wrote the 3rd chapter, then the 4th, the 2nd, and the 1st -- 3,4,2,1. The reason for this is simple. I wrote the chapters of this dissertation by first inquiring about the relevance of an integrative view of evolutionary, cultural, and computational psychiatry. This resulted in chapter 3. Then, I wanted to confirm whether such an integrative model could be implemented with the models of computational psychiatry. This

resulted in chapter 4. Finally, reflecting on my work, I asked whether the integrative view in chapter 3 was philosophically sound. This was the real challenge, which chapters 1 and 2 address. Here, I will also introduce the chapters in the wrong order, to help the reader follow the reasoning behind the evolution of my research project.

Chapter 3, titled *Integrating Evolutionary, Cultural, and Computational Psychiatry: A Multilevel Systemic Approach* contains the negative (i.e., criticism) and positive (proposed alternatives) parts of my argument. Chapter 3 is a targeted literature review that frames the problem that motivates this dissertation; namely, that psychiatry lacks a unified, non-bio-reductionist and principled understanding of mental disorders that connects the many processes making up the mind, from culture to neurocognition and evolution. Describing that problem corresponds to the negative part of my argument. The positive part of my argument proposes the integration of the three principled approaches to psychiatry mentioned above into an Evolutionary, Computational and Cultural model of mental disorder (ECC).

Before we continue, I should offer a word on the nature of the model presented in Chapter 3. I learned during my PhD that the word "model" for philosophers of science is a charged word (as are most words for philosophers). I also learned that there is a rich literature on what a model is (for a review, see Weisberg, 2013). However, I will not endeavour to situate the proposed model in the modelling literature in philosophy. To be frank, beyond the fact that this is not a philosophical question relevant to my project, the reason why I will not try to situate the proposed model in the philosophical modelling literature is that I am not entirely sure how best to characterize the type of model that is at the core of this dissertation. I suspect that it is closer to what George Engel's biopsychosocial model was meant to be (Engel, 1981); that is, a *scientific ontology* that organizes knowledge, and that can be extended to real-world applications in clinical practice.

I am aware, however, that this is not a common way to think about the sort of model I am proposing in chapter 3. The second model that I propose in chapter 4, and that implements the view developed in chapter 3 can be more easily defined. That model is a computational is a computational model based on the theory of active inference (Parr et al., 2022) — something that accounts for the functional relation between some parts that change over time and whose change can be represented numerically, as conforming to some equations, which can be written in the form of an algorithm, the running of which reveals features of the functional relations that would be otherwise hard to envisage. In that sense, the model is a computational model (Weisberg, 2013).

Typically, these models are used to simulate the behaviour of a system, like the brain, and much of the modelling literature in philosophy on active inference is about figuring out what position we should adopt with respect to the relation between models and what it is they model; in particular, whether we should be realist, anti-realist, or pragmatic. But for what matters to this dissertation, the potential applications of the model in psychiatry. we do not need to resolve these problems because the model I propose is more than simply a computational model. Rather, it may be more desirable to read that model as a medical model, much in the way Engel intended it with his biopsychosocial model.

I am also aware that in medicine, there is an equally complex — though more pragmatic — taxonomy of the types of models (Huda, 2019). For example, there are explanatory and aetiological models of diseases which

are used to explain the causes of the affliction and guide research and intervention, and there are models that underlie medical diagnosis, which are used to guide assessment, classification, the management of symptoms and the organization of care. One can also add institutional models of medicine, which define operations at the social and administrative level, thereby linking medicine with other domains of social activity like the law and economy.

For Engel, a medical model (e.g., the biomedical model) is what underwrites all these activities within medicine. Thus, a medical model (i.e., conceptual model) is primarily a scientific ontology. In the terms of the tradition of the sociological and anthropological approach to science (Fleck, 1979; Kuhn, 1962; Latour, 2000; Pickering, 1995), a medical model in Engel's sense would be a thinking pattern born from a collective way of perceiving an object of inquiry and from the delineation of in-groups and out-groups through normative discourse formation, based on the sense of belonging to a scientific community. Such thinking patterns are reproduced through scientific education, rooted in material apparatuses and techniques, and implicitly shape conceptions of truths available to newcomers, and in the case of medicine, styles of clinical intervention. As Engel puts it:

> *"How physicians approach patients and the problems they present is very much influenced by the conceptual models around which their knowledge and experience are organized. Commonly, however, physicians are largely unaware of the power that such models exert on their thought and behaviour. This is because the dominant models are not necessarily made explicit. Rather they become part of the fabric of education that is taken for granted, the cultural background against which they learn to become physicians. Their teachers, mentors, texts, the practices they are encouraged to follow, and even the medical institutions and administrative organizations with which they associate, all reflect the prevailing conceptual models of the era." (Engel, 1981, p. 101)*.

Engel's biopsychosocial model is also a response to the problem of reductionism in biomedicine, which tends to exclude the person in assessments of potential causes of disorders. Engel's biopsychosocial model adopts a systemic view of biology that views traits as existing in a "hierarchically arranged continuum" that goes from single cells to the biosphere (see fig. 1). For Engel, it is within such a model that scientific knowledge must be organized. The model proposed in chapter 3, and which is realized in chapter 4, can also be viewed as implementing the systemic view of biology that underwrites the biopsychosocial model. It should very much be viewed as a dynamic representation of what figure 1 presents (spanning scales or levels from the nervous system to social-cultural systems); dynamics that are tailored to the generation of specific symptoms characteristic of mental disorders. Although, as I said, I do not yet know exactly how to characterize the model on offer in chapter 3, I believe it could be described as (i) reporting an ontology and an epistemology of psychiatry — a conceptual model, that can influence practice and "*become part of the fabric of education that is taken for granted"* (Engel, 1981, p. 101); (ii) as a means of organizing knowledge; and (iii) as a practical guide to orient the way clinicians take care of their clients. The view of the biopsychosocial model as a scientific ontology that guides the formation of other models in medicine is the correct (or perhaps most forgiving) view of the model I propose in chapter 3.

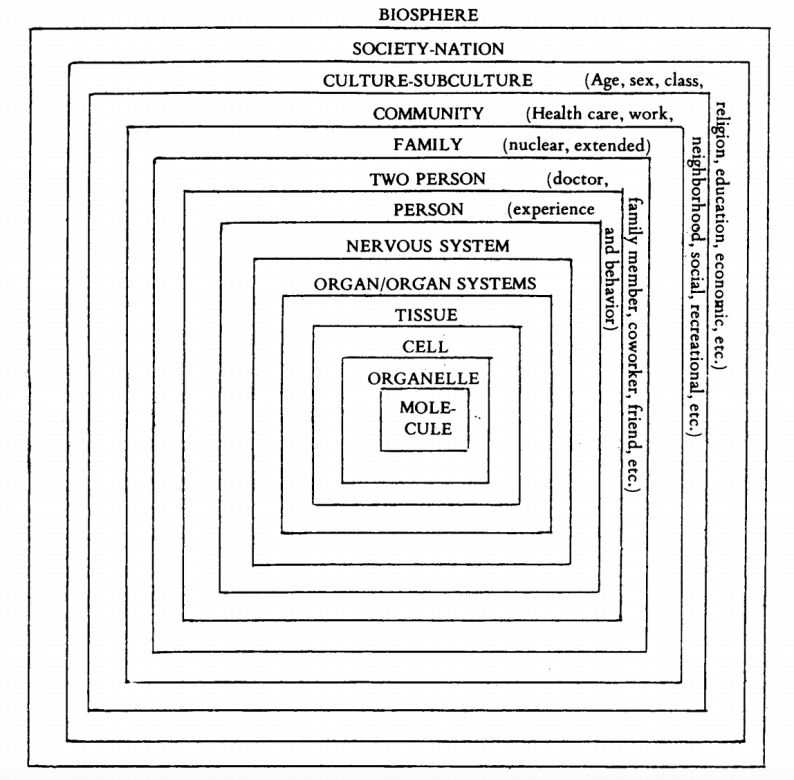


Figure 1. For Engel, each level is a stable configuration that possesses unique dynamics and is part of the dynamics of the whole. The stability at each level is maintained by the relation between internal and external dynamics. Each level possesses boundaries that delineate it as part of the whole, and there is information flow across the entire biosphere (Engel, 1981).

Chapter 4, titled *Why depressed mood is adaptive: A numerical proof of principle for an evolutionary systems theory of depression* applies the integrative evolutionary, cultural and computational approach, developed in Chapter 3, to provide a proof of principle for the Evolutionary System Theory of depression developed by Paul Badcock (Badcock et al., 2017). The Evolutionary System Theory of depression is itself an integrative approach of evolutionary and computational rationales in psychiatry and therefore highly compatible with the model developed in chapter 3.

The proposed computational model in chapter 4 further integrates the sociocultural rationale to design a simulation of the symptoms of major depression. The model allows us to observe, *in silico*, the effects of various types of pharmacological and social interventions on the course of depressive illness. These simulations feed from the more general ontology developed in chapter 3. Thus, chapter 4 is not only a proof of principle of the evolutionary systems theory of depression, but also, in a sense, a proof of principle for the more general 'eco-socio-computational' way of thinking about mental illnesses that I attempt to develop in this dissertation.

I should emphasize that the model of chapter 3 is different in nature than that of chapter 4. As we said, the model of chapter 3 is closer to what Engel had in mind with the biopsychosocial model. In turn, the model of chapter 4 is what people have in mind in computational psychiatry when they develop computational

phenotypes of mental disorders. The model of chapter 3 should be viewed as the theoretical motivation for developing the computational model of chapter 4.

Chapter 2, titled *Extended active inference: Constructing predictive cognition beyond skulls*, is one of the two chapters that clarify and set the theoretical foundations for the model we develop in chapter 3. One of the theoretical foundations I needed was the claim that mind is spatiotemporally distributed, or 'extended', and that such extensions are constitutive of bona fide functions of the mind.

Theoretically and from the point of view of its design, the model implemented in chapter 4 is motivated by the idea that individual minds cut across spatiotemporal scales. This is also the idea that the biopsychosocial model tries to implement, which is represented by figure 1 above. The mind is made of things beyond the brain (spatial scale), and its behaviour is shaped by phenomena that unfold over time scales longer (and shorter) than the lifespan of the individual (temporal scale). This means that interventions on mental illness can work by acting on different subcomponents of the mind, unfolding at multiple spatial and temporal scales.

The interventions that return our simulated agent to normal functioning after an episode of major depression in chapter 4 works precisely in this way. They assume that the phenotype of depression ought to be modelled with parameters that are spatially distributed across internal and external components of the phenotype, and whose dynamics unfold at various temporal scales, from evolutionary (either fixated or more stable over development), to socio-environmental (slow changing parameters) and neurocognitive (fast changing parameters) processes. This modelling assumption had to be supported, theoretically, which again, is what Chapter 2 does for us in this dissertation.

Chapter 1 is titled *The free-energy principle: It's not about what it takes, it's about what took you there*. According to the free-energy principle, any (living) system that exists over time and that can change over time will change to minimize its free-energy, where free-energy is a proxy for the (negative log) evidence (in Bayesian terms) of the model the system entails. The challenge with the free-energy principle is that many things have been said about it; things that left unclear would make my argument untenable. One of those things is the claim according to which minimizing free-energy is a sufficient criterion for defining what counts as a living system. Because I use the free-energy principle to model the behaviour of non-living entities — such desire paths and earthworm burrows, see chapter 2 — someone might think it follows that such entities are in fact alive. Indeed, it is often said in the literature on the free-energy principle that minimizing free-energy is the mark of life and cognition. However, I argue that this is incorrect. The free-energy principle, at best, gives a necessary criterion for life (i.e., every time you will see a living thing, that living thing will be minimizing its free-energy). But the inverse claim is not true (i.e., each time you see something minimizing free energy, this will be a living thing). The failure to recognize that minimizing free-energy is *not* a sufficient condition for life becomes a serious problem when one seeks to apply the free-energy principle to nonliving things. Chapter 1 debunks the sufficiency claim often attributed to the free-energy principle and thus clears the way, pre-emptively for chapter 2, which applies the free-energy principle to nonliving things into which the mind extends.

Chapter 1: The free-energy principle: It's not about what it takes, it's about what took you there

Introduction to chapter 1

The agenda of chapter 1 is twofold. First, the explicit goal is to debunk the idea that minimizing free-energy is a sufficient condition for life; what I will call the 'strong claim'. The second goal of chapter 1 is to give the reader a clear understanding of what free-energy is conceptually and computationally. Chapter 1 provides an example of free-energy minimization that is illustrated by a network that maps out the components involved in free-energy minimization. It will be useful to keep the visualization of the network in mind, as this network will be reproduced throughout the dissertation to unpack different dynamics.

By debunking the strong claim, this chapter prevents a serious problem that one might think follows from the modelling strategy I employ in chapter 2, and which will be used — and referred to — in the rest of this thesis. The problem is that given that it is often assumed that minimizing free-energy is a sufficient condition for life and cognition and given that I will assume that free-energy minimization occurs in non-living entities, one might think that my position is that nonliving parts of the biosphere are alive in some sense. This is not the line of argument I want to pursue. I am not advocating for panpsychism or “pansentiencism”, nor do I think that inanimate things should be treated as living things.

1 Introduction

Sometimes, arguments in the literature on the free-energy Principle (henceforth FEP) give the impression that in order to be alive, viz. to count as a living system, one must minimize free energy. Such a claim does not straightforwardly apply to the free-energy principle, however, and this is what this chapter will demonstrate. Minimizing free-energy does not entail life. Rather, the argument is that if you are alive, it probably means that you have done something like minimizing your free energy, which is the (Bayes) optimal thing to do when your life depends upon solving complex inference problems. This is a subtle, but crucial point to getting the story straight. I shall call this the ‘entailment problem’; that is, the confusion in the entailment relation between

free-energy minimization and life. Here, the notion of entailment refers to the implication (i.e., first order logical property) between free-energy minimization and the fact of "displaying some life-related processes".

The entailment problem, it seems to me, stems from the fact that there are at least two types of claims one can conceive of when thinking about the relation between life and free-energy minimization. Or rather, about *survival*, and free-energy minimization; although, under the FEP, these appear to be synonymous. Minimizing free-energy is the process whereby one maintains one's structural integrity in the face of environmental perturbations by revisiting one's most probable organization of physiological states (Friston, 2013; Kirchhoff, 2015). It is in that sense that minimizing free-energy is considered a condition for life. One can equate 'survival' with 'life', since one supposes the other under the FEP; 'if I survived, it means that I maintained my structural integrity in the face of environmental perturbations'; 'maintaining my structural integrity is what qualifies me as living.'

Now, it might be said that metamorphic organisms, despite not keeping their structural integrity, should be considered as living organisms. This was noted by Kirchhoff et al., 2018 and Clark, 2017 and others (Friston and Stephan, 2007; Andrews, 2022; Sims, 2021). From the point of view of the FEP, when considering such metamorphic organisms, it may be said that it is the lifecycle that corresponds to the thing whose integrity is maintained over time (e.g., over evolutionary time), not the specific form that the system takes at one stage of its development (e.g., the adult form of a frog). This casts the FEP within the realm of the process ontology (vs. substance ontology). Similarly, while it may be argued that life is a state instantiated by an organism at time 't' whereas survival is a process with duration (e.g., the endurance of life from t to t+1), under the FEP (from the perspective of process ontology), it is unclear whether these two notions — life and survival — really have a different referent. It might be argued that both life and survival both refer to an enduring process that advocates of process ontology would call 'organism' (cf. Dupré, 2020). While there is certainly a fuller discussion to be had concerning the correct unit of analysis for free-energy minimizing organisms and the meaning of life and survival under that theory, a *critical* discussion of these issues is beyond the scope of this chapter.

The first type of claim on the relationship between life and free-energy minimization is a strong type according to which minimizing free-energy is a sufficient condition for life. This claim has been called the overly generous claim (Kirchhoff & Froese, 2017). Such a claim is attractive since it suggests that knowing what is involved in minimizing free-energy (e.g., possessing a Markov blanket) will inform us about what it takes to be alive. Such a strong claim would allow us to generalize the scope of the FEP to the full range of possible beings, and in so doing, it would allow us to predict which of those will pass the bar for qualifying as 'living'; it would allow one to identify 'what it takes' to be alive from the point of view of the FEP.

The second type of claim is a weak type according to which if a system is currently alive, it means that it minimized its free energy. Such a type of claim does not assume that the FEP is designed to set the bar for the sufficient conditions for life or meant to predict what things may or may not be alive. Rather, it limits the scope of application of the principle to beings that we think are alive, now, and enables us to know the necessary conditions under which those beings can be living – i.e., can actively resist the loss of structural integrity; 'what took them there?'

In the primary literature on the FEP, we can often read passages that may be interpreted as making strong claims, such as "minimization of free-energy may be a necessary, if not sufficient, characteristic of evolutionary successful systems" (Friston & Stephan, 2007, p. 26), and "systems that do not minimize free-energy cannot exist" (Friston, 2013, p.2). And so, people have reacted saying things of the sort "the right direction of explanation must go from minimizing free-energy to survival. Yet insofar as FEP implies a causal story about that direction of explanation, it appears to be wrong. On the one hand, minimizing free-energy cannot be sufficient for survival" (Klein, 2018, p. 12). Here, Klein advocates the impossibility of a strong claim in favour of the FEP. In the secondary literature, claims such as the aforementioned ones found in the primary literature have led some people to claim that the goal of the FEP is to discover the necessary characteristics of living systems, and that the free-energy minimization is an 'imperative' of life (Van Es, 2020). Here, one might argue that the terms 'imperative' and 'necessary' are correctly employed in the weak sense – i.e., in the sense of 'if life, then free-energy minimization has occurred' – but not a sufficient one, in the sense of 'if free-energy minimization occurs, then life follows. But had the relation between life and free-energy minimization been correctly interpreted as merely necessary, some of the problems that van Es' claim is meant to motivate would simply not apply. Indeed, although it is hard to find direct evidence of what I called the entailment problem in the literature, that problem often transpires through some of the challenges that motivate philosophers to write on the FEP.

Take for instance the problem of scope, which is considered a serious problem among others by van Es. The scope problem refers to the danger of being over generous with applications of the FEP, out of fear of being overly generous with what we count as living (or as having a mind). Obviously, this is only a problem for someone who thinks that the FEP is meant to provide sufficient conditions for life (or mind). For instance, referring to a passage of Karl Friston's seminal paper 'Life as we know it' (2013), Kirchhoff and Froese (2017) say that:

> "*Strictly speaking, what Friston says here is that for any system to exist it must work to minimize free energy. This commits Friston to one of the following three implications. First, if free-energy minimization is sufficient for mentality, then every system has a mind, even if not all systems are alive. Second, if free-energy minimization is enough for life and mind, then all systems that exist are both alive and mental. Finally, biological systems, like all other existing systems, need to work to minimize free energy. The last option states that free-energy minimization is not a property of only living systems, and as such sets up one of the two following implications. Either (option one) the FEP places mentality in a class of systems that includes but is not limited to living systems, and therefore veers towards some form of panpsychism. Or (option two) the FEP equates life-mind continuity with a view that sees life and mind nearly everywhere. […] . Our point is: given that the core concepts of non-cognitivist FEP—approximate Bayesian inference, ergodicity, Markov blankets and so on—can be applied to living and cognitive systems, on the one hand, and seemingly non-living and non-cognitive systems, on the other, there is a clear danger of these concepts being over-broad in their application, resulting in either seeing life and mind nearly everywhere or in the FEP lacking explanatory power when having to address the nature of life and mind and their relation to one another*" (Kirchhoff & Froese, 2017, p. 10-11).

There is no such danger associated with the FEP for the simple reason that it is not because a system minimizes its free-energy (and that system has a Markov Blanket) that that system is alive. Again, free-energy

minimization is not a sufficient condition for life (or mind). It seems to me that the problem of scope would only worry those who believe that the FEP makes a strong, sufficiency claim about the relation between life (or mind) and free-energy minimization.

Other standard manifestations of what I call the entailment problem take the form of a critique of the 'testability' and 'tautology' of the FEP, which would be worries for the strong claimers, and for people who are generally worried about the explanatory power of the FEP, as mentioned by Kirchhoff and Froese. I do not have the space to elaborate on this here, plus this has already been done (Colombo & Wright, 2018). Instead, in this chapter, I simply dissolve what I called the entailment problem by providing a numerical example of free-energy minimization in a hypothetical organism. I conclude with some brief epistemological remarks that may be of interest for those who worry about the explanatory power of the FEP.

The proposed numerical example will clearly demonstrate why minimizing free-energy can generate both Bayesian adaptive and Bayesian maladaptive behaviour, leading to survival, or death, accordingly. The proposed numerical example demonstrates that minimizing free-energy is not sufficient for life – the strong claim. The proposed numerical example demonstrates the necessity claim; the idea that under the right conditions, remaining alive means that free-energy was minimized – the weak claim. The relevance of the weak claim should become apparent through the reading of the numerical example, which will show that under the right conditions, minimizing free-energy should allow the maintenance of structural integrity. Hopefully, this numerical example will appease those who want to raise worries, implicitly or explicitly, about the — non-existent — FEP strong claim, or about the apparently less interesting weak claim.

2 Minimizing free energy: for better or worse

2.1 Some conceptual distinctions between Bayes and the free-energy principle

Bayesian approaches to animal behaviour propose that one can model organisms as representing their relation to environmental states using priors and a likelihood (McNamara et al., 2006). Let's call those representations Bayesian 'beliefs'. On the basis of those beliefs, organisms generate adaptive behaviour. Bayesian beliefs represent (i) the probability of environmental states, prior to observing an environmental signal (a.k.a. prior); and (ii) the relation between environmental states and observed environmental signals (a.k.a. likelihood). Bayes theorem, from which terms such as prior and likelihood come from, is typically expressed as an evidentiary relationship between some prior hypotheses (P(H)) and the observation at hand, or data (E): $P(H|E) = [P(E|H)P(H)]/P(E)$.

The free-energy principle is a Bayesian formulation of the manner in which organisms infer the posterior probability of their prior beliefs after having observed an environmental signal with a given likelihood, and in so doing infer some hidden, or unobserved variables. What stands for the 'H' are the unobserved variables

whose prior probability P(H) forms the hypothesis, and what stands for the 'E' are the sensory signals organisms receive (the data). Hence, it is often said that under the free-energy principle, organisms are viewed as embodying a 'hypothesis', a 'belief' or a 'best guess' about the cause of their sensations, or sensory signals they receive (Allen & Friston, 2016; Bruineberg & Rietveld, 2014; Friston, 2011).

Under the FEP, the evidentiary relation explains the manner in which organisms self-evidence (Hohwy, 2016), where the 'self' means evidencing beliefs about oneself in the world. Because beliefs are embodied by the organism, and are thus the organism's own states, the uncertainty in the likelihood and the prior can be viewed as representing the uncertainty inherent to the biological apparatus (e.g., noise in the signal transmission across the nervous system), instead of the uncertainty of the world (e.g., fluctuations in states of the world generating the signals), as would be the case under typical Bayesian models. Under the FEP, uncertainty should thus be read as reporting a Bayesian 'credence score' over the organism's own beliefs, as it reports the probability of a state or hypothesis relative to other possible hypotheses. In the case of the likelihood, the credence score is over sensory beliefs relative to states (e.g., 'is this more probably warm or more probably hot?'). In the case of the prior, the credence score is over the hypotheses the organism entertains prior to sensing the water temperature (e.g., 'am I probably in the ocean or in my bath?'). Of course, these beliefs, hypotheses, or best guesses are implicit and subpersonal, as they are meant to be realized by the organism's (neuro)physiology. This begs another important question: are priors subjective or objective under the FEP?

Initial priors can be of two kinds: (i) objective, or (ii) subjective. Objective priors are typically based on frequencies (e.g., priors that report distributions based on empirical data). When the frequencies are unknown, an equiprobable (flat) prior should be favoured. Objective priors thus conform to some rational constraints beyond Bayesian rationality. Subjective priors, in turn, refer to the psychological dispositions of the system of interest, or to the person specifying the system of interest (e.g., priors that report propositional attitudes). Subjective priors do not need to conform to constraints of rationality. The simple answer to the question of whether priors are objective or subjective under the FEP is that they are subjective. As we said above, they track the confidence of a system's beliefs. We could thus carry on with that in mind.

However, there is an interesting detail on that question that may be worth mentioning. Under the FEP, priors do not conform to rationality beyond the rationality of the inference *per se*, but nor are they rationally unconstrained. There is rationality beyond Bayesian rationality that comes from the way variational Bayes is embodied by the system (Hohwy, 2020). As we will see in detail later on, the inference of the posterior distribution requires finding an approximation to that posterior (denoted as 'Q' later on), which then becomes the prior used in the next cycle of inference. That approximate posterior determines what is embodied by the organism. The update having led to the subjective posterior at time t+1 operates by finding the subjective posterior that would *best* approximate the true subjective posterior distribution. The meaning of 'best' just is being close to 0 free energy. That true subjective posterior is that which one would find with exact Bayesian inference – more on this later, and crucially never exists. Hence it is sometimes said that it forms only a reference point to perform the inference (Ramstead, Kirchhoff and Friston, 2019).

The rational constraint over the priors is the fact that the approximate subjective posterior 'Q' (or future prior) will not only be Bayesian, but also will always be the 'best guess' relative to what the true posterior *ought to be*. In short, under the FEP, even though priors refer to psychological states of the system, updates of the system make those priors an approximation of what they 'should' have been, had the prior been updated with exact Bayes. Thus, it might be said that priors under the FEP cut across the objective / subjective dichotomy. They are subjective while satisfying a rational constraint mandated by the existence of the system *per se*.

2.2 The numerical example

The numerical examples below operating under the following scenario (see fig. 1). Consider an organism that infers whether an external event A or B took place. For the organism, A and B are part of the class R and form the representations by the organism of the external events represented by A or B. A and B are inferred when receiving a chemical signal part of the class S, which can be alpha or beta.

We assume that before observing any signal, the probability of A is p, and the probability of B is 1-p. Given the environment in which the organism finds itself, the probability of observing a signal alpha under A is m, and the probability of observing beta under A is 1- m. We assume that p is equal to .8, and that m is equal to .7. The opposite applies to B. We stipulate for the sake of the numerical example that representing A when receiving alpha, or B when receiving beta leads to survival, and that the opposite leads to death. Because inference is biophysically realized, representing, or inferring A or B could also be interpreted as producing a metabolic response (not necessarily an action) to alpha or beta. Heuristically, the reader can assume that we simply stipulate that inferring B when sensing alpha or A when sensing beta is a maladaptive metabolic response that prevents from maintaining structural integrity. The prior probability P(R) and the likelihood P(S|R) can be visualized as follows:

$$P(A) = p = .8$$
$$P(B) = 1 - p = .2$$

$$\textbf{Prior: } P(R) = \begin{bmatrix} .8 \\ .2 \end{bmatrix}$$

$$P(\alpha \mid A) = m = .7 \qquad (1)$$
$$P(\beta \mid A) = 1 - m = .3$$
$$P(\alpha \mid B) = 1 - m = .3$$
$$P(\beta \mid B) = m = .7$$

$$\textbf{Likelihood: } P(S \mid R) = \begin{bmatrix} .7 & .3 \\ .3 & .7 \end{bmatrix}$$

Assuming that the internal computation that our organism performs conforms to Bayes theorem (McNamara et al., 2006; Okasha, 2013), computing the posterior probability of A or B relative to the environmental signal *and representing the most likely state amounts to selecting the state with the highest posterior probability.* Let's infer the posterior probability of A after observing, say, alpha. To do this, we would apply Bayes theorem as follows:

$$P(A) = .8$$
$$P(\alpha \mid A) = .7$$
$$P(B) = .2$$
$$P(\alpha \mid B) = .3$$

$$\underbrace{P(A \mid \alpha) = \frac{P(\alpha, A)}{P(\alpha)}}_{\textbf{Bayes rule}}$$

$$\underbrace{P(\alpha, A) \; = P(A)P(\alpha \mid A) = \; .8 * .7 = .56}_{\textbf{Joint probability of } A \textbf{ and } \alpha}$$

$$\underbrace{P(\alpha) = P(A)P(\alpha \mid A) + P(B)P(\alpha \mid B) = .56 + .06 = .62}_{\textbf{Marginal distribution (a.k.a. model evidence)}}$$

$$P(A \mid \alpha) = \frac{P(\alpha, A)}{P(\alpha)} = \frac{.56}{.62} = .9032 \qquad (2)$$

Eq. 2 takes the prior probability of A, which is .8, and multiplies it by the likelihood of A under signal alpha, which is .7, in order to get the joint probability of A and alpha, which is .56. In order to find the posterior probability, one must divide this joint probability by the marginal distribution, which is simply the sum of the joint probability for A and B under signal alpha, respectively; or alternatively, the prior for B times the likelihood for B under alpha, plus the prior for A times the likelihood for A under alpha. Exact Bayesian inference yields a posterior probability of .9032 for state A after having observed the signal alpha (and a posterior of .0968 for B, since the posterior distribution must sum to 1). This means that after seeing alpha, an exact Bayesian organism would have represented A with ~90% confidence, and thus would have survived.

With exact Bayesian inference, one uses the marginal distribution to find the posterior probability. This assumes that the organism could sum over the probability of outcomes under both A and B. However, it is unclear whether living systems have sufficient computational power to accomplish that (Bogacz, 2017; Friston, 2009). For instance, following our numerical example, the signal alpha might have been caused by environmental states A,B,C,…, each of which would have an analogue internal state A,B,C represented by the organism. Thus, the likelihood modelled by the organism might look like this:

$$\underbrace{P(S \mid R)}_{\text{Likelihood}} = \begin{bmatrix} P(\alpha \mid A) & P(\alpha \mid B) & P(\alpha \mid C) & \ldots \\ P(\beta \mid A) & P(\beta \mid B) & P(\beta \mid C) & \ldots \end{bmatrix} (3)$$

Under exact Bayesian inference, all the probabilities in eq. 3, for all states under the observation of interest (e.g., alpha) should be summed over. Doing this will often be computationally intractable, as the organism will entertain multiple different causal representations (e.g., A,B,C…) for the same observation (e.g., a red sensation that might have been caused by a red 'shoe', red 'car', red 'traffic light', red …). This problem underwrites what is referred to in the literature on the free-energy principle and predictive processing as the black box problem (Clark, 2013), the solipsism problem (Hohwy, 2016), or the seclusion problem (Wiese & Metzinger, 2017).

In order to bypass this problem, the FEP models the inference process (e.g., of A or B) performed by organisms as approximate Bayesian inference. Approximate Bayesian inference bypasses the direct evaluation of the likelihood and the marginal distribution when inferring the posterior probability. Note that in biology, similar methods became popular through work in population genetics on the genealogy of DNA sequences (Sunnåker et al., 2013; Tavaré et al., 1997). The central claim of the FEP is that changes leading to behavioural and (neuro)physiological responses in living systems conform to a form of approximate Bayesian inference known as variational Bayes (Beal, 2003; Friston, 2005, 2013; Parr & Friston, 2018).

Now, building on the numerical example above, the following numerical example shows that one can infer the posterior probability for A by minimizing free energy; and with the same inference process and the same likelihood, one can find a posterior that gives high confidence to B. Given that representing A when observing alpha leads to survival, and representing B when observing alpha leads to death, the following numerical example will demonstrate that minimizing free-energy is not a sufficient condition for life, as it can lead to the exact opposite -- death.

Note that the scope of the following numerical example is deliberately limited. The goal is to demonstrate that minimizing free-energy can lead to maladaptive inference when performed with the wrong priors, all things being kept fixed. If the priors are allowed to update, the inference should lead to adaptive behaviour. This is an important point to which we will come back below. Adaptivity is guaranteed by the extent to which the priors match the environmental constraints, more than by the nature of the machinery employed to perform the inference (e.g., free-energy minimization or exact Bayes). That being said, the machinery that allows the inference will play an important role in allowing priors to match environmental constraints. The following example of free-energy minimization is provided in the sole purpose of supporting our response to the entailment problem. The goal is to give a formal intuition as to why minimizing free-energy is not sufficient for life understood as the preservation of structural integrity. By no means should the following numerical example be viewed as an exemplar of the manner in which free-energy minimization operates, mathematically. The following numerical example simply illustrates the concepts engaged in this chapter and does not provide a complete understanding of the mathematical apparatus of the FEP. Technical readers should refer to Buckley et al. (2017) and Bogacz (2017), or Smith et al. (2021).

Free-energy 'F' is defined as follows:

$$F = -\sum_{R} Q(R)\left[\ln\frac{P(R,\alpha)}{Q(R)}\right] \quad (4)$$

Eq.4 says that free-energy on the left side of the equation is equal to the (negative) sum of the log ratio of an approximation to the posterior for A and B (Q(R)) and the joint probability of those states and signal 'alpha' (P(R, alpha)), multiplied by the approximate posterior (Q(R)). Minimizing free energy, from the perspective of eq. 4, just means finding the approximate posterior Q(R) that will yield the F that is the closest to 0 on the left side of the equation. Q(R) corresponds to the *proposal, recognition, or approximate posterior density* sometimes referred to in the literature on the FEP. It is that Q(R) that is embodied by the organisms -- not to confuse with the P(R,alpha), which would be the joint distribution, or generative model (Ramstead et al., 2019).

Above, using exact Bayesian inference, we had to divide the joint probability of A and alpha by the marginal distribution. Recall that here, we want to remain agnostic concerning the marginal distributions to which we do not have access. We can find the posterior under such constraints by asking, 'what approximate posterior Q(R) gives me the least F'? The answer to that question is the approximate posterior Q(R) that will be the closest to the true posterior.

We know from exact Bayes that the true posterior probability of A given alpha (P(A | alpha)) is .9032, meaning that after observing alpha, our exact Bayesian organism represented state A with ~90% confidence. Now, let's assume that our organism operates under variational Bayes, and that it indeed represented A with the same level of confidence. What would have been its free energy? This can be computed as follows:

$$F = -\sum_{R} Q(R)\left[\ln\frac{P(\alpha,R)}{Q(R)}\right]$$

$$= \left(-\left(\underbrace{.9032}_{Q(A)} * \ln\frac{\overbrace{.56}^{P(\alpha,A)}}{\underbrace{.9032}_{Q(A)}}\right)\right) + \left(-\left(\underbrace{.0968}_{Q(B)} * \ln\frac{\overbrace{.06}^{P(\alpha,B)}}{\underbrace{.0968}_{Q(B)}}\right)\right)$$

$$= .4780 \quad (5)$$

Eq. 5 tells us that the free-energy of an organism with an approximate posterior equal to the true posterior would be .4780; or put another way, minimizing free-energy down to .4780 means representing A with a level of confidence of ~90%. Now let's imagine an organism that would have inferred P(A | alpha) with a probability of .0968, which we know is far from the true posterior:

$$F = -\sum_{R} Q(R)\left[\ln\frac{P(\alpha,R)}{Q(R)}\right]$$

$$= \left(-\left(.0968 * \ln\frac{.56}{.0968}\right)\right) + \left(-\left(.9032 * \ln\frac{.06}{.9032}\right)\right)$$

$$= 2.2792 \qquad (6)$$

Eq. 6 tells us that an organism that would have represented B with ~90% confidence after seeing alpha would have had a free-energy of 2.2792, which is higher than .4780. Based on the current scenario (i.e., B when receiving alpha leading to death), the organism with the higher free-energy would have died. Hence, one might be tempted to agree with the claim that minimizing free-energy is sufficient, if not necessary for survival. Indeed, when comparing eqs. 5 and 6, minimizing free-energy – i.e., finding the approximate posterior that yields the free-energy closest to 0 – guarantees survival, whereby the opposite guaranteed death.

However, minimizing free-energy leads to survival only under the right conditions, that is, if the organism has the right prior beliefs, and the right joint probability, accordingly. Let's imagine the same scenario, with the same likelihood and success conditions, but with inverted prior beliefs. This is conceivable, for instance, if an organism inherits maladaptive prior beliefs (Richerson, 2018). Let's imagine that our organism has inherited a maladaptive, inverted prior:

$$P(A) = .2$$
$$P(B) = .8$$
$$P(\alpha \mid A) = .7$$
$$P(\alpha \mid B) = .3$$

$$P(\alpha, A) = .2 * .7 = .14 \qquad (7)$$
$$P(\alpha, B) = .8 * .3 = .24$$

Posterior of A $= .3684$
Posterior of B $= .6316$

Eq 7. Simply inverts the prior probability we started with in eq.1. and shows the consequence for exact Bayesian inference. With the same likelihood, but an inverted prior, an exact Bayesian organism would have represented state B with ~.63% confidence after seeing alpha; and thus, would have died. As you might suspect it, the same applies to a free-energy minimizing organisms:

$$F = -\sum_R Q(R)\left[\ln\frac{P(\alpha,R)}{Q(R)}\right] \qquad F = -\sum_R Q(R)\left[\ln\frac{P(\alpha,R)}{Q(R)}\right]$$

$$= \left(-\left(.3684 * \ln\frac{.14}{.3684}\right)\right) + \left(-\left(.6316 * \ln\frac{.24}{.6316}\right)\right) \leq \quad = \left(-\left(.6316 * \ln\frac{.14}{.6316}\right)\right) + \left(-\left(.3684 * \ln\frac{.24}{.3684}\right)\right)$$

$$= \underbrace{.9676}_{\text{F when representing B after sensing } \alpha} \qquad = \underbrace{1.1094}_{\text{F when representing } A \text{ after sensing } \alpha}$$

(8)

Eq. 8 tells us that when observing alpha, representing B with ~.63% confidence yields a free-energy of .9676, which is closer to 0 than 1.1094. This means that an organism minimizing its free-energy would have represented B instead of A when observing alpha. In the current scenario, this is fatal. Hence minimizing free-energy *per se* does not entail life; not under the wrong prior. In fact, it can perfectly well entail the exact opposite. And so, it should be clear that claims according to which free-energy minimization provide sufficient conditions for life should not be interpreted as such. There is no logical consequence that goes from minimizing free-energy to life understood as maintaining one's structural integrity.

Generative model (a.k.a. joint probability) $= P(R,S) = P(R)P(S \mid R)$

Prior $= P(R) = \begin{bmatrix} .8 \\ .2 \end{bmatrix} \begin{matrix} A \\ B \end{matrix}$

Likelihood $= P(S \mid R) = \begin{matrix} \end{matrix} \begin{matrix} A & B \end{matrix} \\ \begin{bmatrix} .7 & .3 \\ .3 & .7 \end{bmatrix} \begin{matrix} \alpha \\ \beta \end{matrix}$

Bayesian organism : $P(A \mid \alpha) = \dfrac{P(A)P(\alpha \mid A)}{P(A)P(\alpha \mid A) + P(B)P(\alpha \mid B)}$

Variational Bayesian organism : $Q(A) = \arg\min_Q F \Rightarrow Q(A) \approx P(A \mid \alpha)$

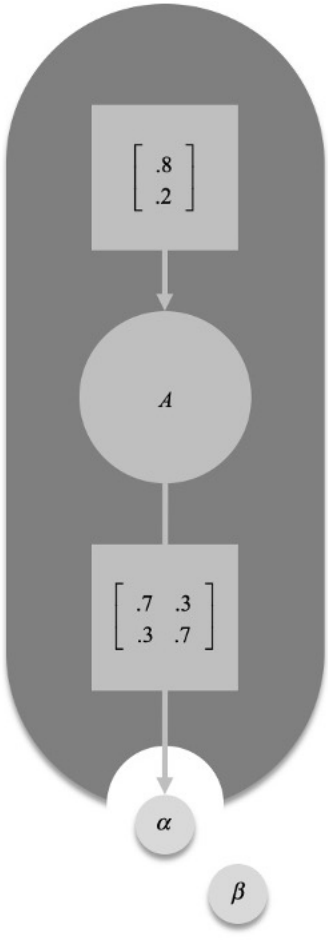


Figure 1. Right panel.: Visual representation of an organism inferring the cause of the observation that it makes. The ensuing beliefs are assumed to complement the external cause of the observation. Here, the

organism observes the outcome 'alpha', and on the basis of its prior and likelihood (i.e., sensory beliefs) finds the posterior value of its beliefs. Given that behaviour is formally equivalent to inference in our simple organism, inferring 'A' as the right beliefs about the most probable cause of observation means biophysically representing 'A'. There is no action involved in our example. The likelihood and the prior are assumed to 'map', heuristically, onto the physiology of our organisms – the prior being some sort of storage of knowledge, and the likelihood being the sensory belief. In more biologically realistic descriptions of behaviour, which require a discussion of active inference, behaviour is the result of a different inference process – that of an action policy (under discrete models). This involves more priors, namely, about the transition between hidden states and often about preferred sensory outcomes. Action then is distinguished from inferring hidden states. It is about inferring another hidden variable, which is the policy. Left panel: The first line represents the organism, formally, as a joint distribution obtained by multiplying the prior and the likelihood (which is biophysically implemented). This joint distribution can also be viewed as a 'generative model', or model of the manner in which sensations are caused by external states. Inferring the posterior probability, based on that joint distribution or generative model, allows the organism to respond adaptively and to generate for itself the right sensation. Indeed, one must distinguish the sensory input (e.g., alpha or beta) from the generated sensation by the organism. The second and third lines represent the prior and the likelihood, formally. The fourth line represents the possible Bayesian algorithm that could be used. The fifth line presents variational free-energy minimization that selects the approximate posterior density (Q(A)).

2.3 free-energy on a wing and a prior?

Although free-energy minimization is not sufficient for life, the numerical example above suggests that there might be an entailment relation that goes the other way around: if you are alive, it might very well because you did something like minimizing free energy. That entailment relation is that which corresponds to the weak version of the life-free-energy entailment relation. Indeed, in our numerical example (eq. 5 and 6), minimizing free-energy led to survival under the right (prior) conditions; and it seems fair to assume that from a Bayesian point of view, minimizing free-energy (or performing a similar form of approximate inference) is what organisms do. This makes the FEP an interesting epistemic principle for researchers interested in development. In a reverse engineering fashion, if we observe a free-energy minimizing organism that is still living at the time that we observe it, we can trust that it has good enough priors to remain alive; and if we observe that that organism behaves maladaptively, we have good reasons to doubt the viability of its current priors. The goodness of priors, of course, rests on the extent to which priors match the sort of challenges the organism is currently exposed to (e.g., if you represent 'B' when sensing 'alpha', you die, and so a good prior is a prior that makes you represent A more often than not – has higher credence on A). The consequence of this is that one can bring the sufficiency claim back into the game if one assumes that the organism is endowed with adaptive prior beliefs.

When equipped with the right prior beliefs (i.e., the priors that represents accurately the states that are conducive to survival), one can get closer to the idea that minimizing free-energy may be sufficient for life; i.e., that it is all you need to be qualified as living, or as maintaining your structural integrity under the free-energy principle. The point here is that the choice of prior, whether under Bayesian or approximate Bayesian regimes is the real concern, since the entailment relation between life and the FEP entirely depends on the adaptivity of those priors. Assuming that priors are genetically inherited, the entailment relation between the FEP and

life will be predicated on evolutionary processes. Interestingly, some have argued that the adaptivity of priors can also be guaranteed by free-energy minimization operating at the population level, as a form of natural selection (Badcock et al., 2019; Constant, Ramstead, et al., 2018; Friston, 2010, 2013; Friston & Stephan, 2007; Hesp et al., 2019; Ramstead et al., 2017; Sella & Hirsh, 2005). To make sense of this, simply imagine that instead of modelling an organism with states A and B minimizing free energy, we are modelling a population with different genotypic states AA Aa aa, each having a prior probability, and each being more or less likely under observable environmental patches. Minimizing free-energy at the population level would allow natural selection to converge on the Bayesian gene pool distribution, that is, the approximate posterior distribution for genotypes that is the closest to the true posterior distribution under reproductive observations. This means that inherited genotypic priors sampled from the approximate posterior distribution at the genotypic level should be well tuned to the environmental pressures that have caused the reproductive success (i.e., observations). By extension, individuals having received the most probable genotype will have genotypic priors that provide the right prior conditions for successful behaviour (e.g., representing A when observing alpha).

However, even such a multiscale free-energy minimization rationale does not guarantee that organisms with the right inherited priors won't undergo somatic mutations, or simply neural lesions that would change the distribution of inherited priors, therefore biasing free-energy minimization over development towards faulty inference, death and the inability to maintain structural integrity.

3 Future direction: free-energy minimization as a historical scientific principle?

The dissolution of the entailment problem puts us in a good position to move on to another related difficulty in the philosophical literature on the FEP, which is, this time, of an exegetical kind. If minimizing free-energy is not sufficient for life or survival, how should we interpret statements such as "the minimization of free-energy may be a necessary, if not sufficient, characteristic of evolutionary successful systems" (Friston & Stephan, 2007, p.428)?. I conclude with an epistemological remark on the meaning of that statement.

The FEP on its own is a principle, namely, a foundation for reasoning about things (e.g., living things). In this chapter, we approached the FEP as such. However, the FEP can also be read more broadly as a research program that uses FEP reasoning patterns to generate scientific hypotheses. This involves implementing FEP reasoning into a theory called active inference, which is routinely used to study various cognitive functions (for a review see Da Costa et al., 2020). As a research program, the FEP can be used to generate statements that are normative in the strong sense. Such statements can be tested using scientific standards for hypothesis testing (Smith et al., 2020; 2021).

As a reasoning pattern, the FEP can be used to generate postdictive statements (cf. Friston et al., 2017)[1]. Accordingly, FEP reasoning might be interpreted as a principle akin to those found in postdictive sciences (a.k.a. historical sciences) like geology, palaeontology, archaeology, or any science that deals with irreproducible causes (Cleland, 2002). Postdictive scientific statements are concerned with what 'must have been the case', instead of 'what will be' the case. A statement such as "the minimization of free-energy may be a necessary, if not sufficient, characteristic of evolutionary successful systems" is probably such a postdictive statement. That statement should be interpreted as claiming that free-energy minimization must have occurred if a system is evolutionarily successful – not the other way around. Nonetheless, this is an interesting statement because if free-energy has occurred, the system in question can be modelled as if it possesses the features allowing for free-energy minimization (e.g., a Markov Blanket). One can then start inquiring about whether those features help us understand the sort of dynamics implemented by the (neuro)physiology of the system, in a predictive fashion (e.g., with the FEP as a research program). Hence, it is sometimes said that the FEP, as a foundation for reasoning, is a 'guide to discovery' (Ramstead et al., 2017).

According to Cleland (2002), historical scientific methodology enables scientists to generate historical hypotheses about the best causal explanation for some observations, based on the accumulation of evidence about the causal structure that might have led to those observations (e.g., evidencing the asteroid-impact hypothesis of dinosaurs' extinction using fossil records of asteroid's impact). In historical sciences, an 'investigator' starts by observing some puzzling traces, or the effects of a cause in the distant or proximal past. The investigator then postulates some hypotheses about the cause of the observed effects. Testing a historical hypothesis then just means accumulating more traces to evidence one of the competing historical hypotheses. These new traces are 'smoking guns', which are meant to shift the 'balance of probability' towards one of the competing hypotheses. A historical hypothesis is defined by the pattern whereby it is evidenced and by its ability to account for those smoking guns with a unifying and compelling causal story.

FEP reasoning yields historical hypotheses because it operates a historical evidentiary pattern and provides a compelling unifying causal story. It operates a curious evidentiary pattern, though, because it assumes that both the investigator and the thing under investigation conform to that evidentiary pattern. That pattern is free-energy minimization, *per se*. For instance, for the organism in our numerical example, the hypotheses were A or B. The smoking guns were the sensory observations 'alpha' or 'beta'. The (self)evidencing activity whereby the organism 'tested' those hypotheses were biophysically realized variational Bayes using the sensory observation to evidence the hypotheses about itself (e.g., A; B). Then, as a person who used the free-energy principle in the numerical example above, the puzzling trace for which I was seeking a causal explanation was the survival of the organism. That was my observation. The causal story or hypothesis for that observation under the conditions we imposed to our simulated organism was the free-energy principle, the inference over which led me to write the chapter you are reading at the moment. That chapter functioned as sensory evidence for my hypothesis (e.g., when writing down the number and seeing they were adding up).

[1] It is important to note that the FEP includes processes other than free-energy minimisation. It also includes expected free-energy minimisation (and generalised free-energy minimisation, (Parr & Friston, 2019)). While minimising free-energy endows the organism with postdictive inference, minimising expected free-energy endows the organism with predictive inference. This is due to the simple reason that the outcomes and states involved in the inference process under expected free-energy minimisation are in the future, not the present. Effectively, this means that inferring one's beliefs about states of the world means inferring what will most likely be seen under those beliefs, and under a given sequence of action to be engaged (i.e., action policy).

And that chapter is the observation that you are using to evidence your hypotheses concerning the claim I set at the start of the chapter, namely, that free-energy minimization is not sufficient for life. Fidel to the unifying grip of hypotheses in historical sciences, the free-energy principle is meant to account for all of that – you, me and the organism under study, in a unifying fashion.

References

Allen, M., & Friston, K. J. (2016). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese*, 1–24.

Badcock, P. B., Davey, C. G., Whittle, S., Allen, N. B., & Friston, K. J. (2017). The Depressed Brain: An Evolutionary Systems Theory. *Trends in Cognitive Sciences*, *21*(3), 182–194.

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. University of London London.

Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *Journal of Mathematical Psychology*, *76*(Pt B), 198–211.

Buckley, C. L., Kim, C. S., McGregor, S., & Seth, A. K. (2017). The free-energy principle for action and perception: A mathematical review. Journal of Mathematical Psychology, 81(Supplement C), 55–79.

Bruineberg, J., & Rietveld, E. (2014). Self-organization, free-energy minimization, and optimal grip on a field of affordances. *Frontiers in Human Neuroscience*, *8*, 599.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioural and Brain Sciences*, *36*(03), 181–204.

Clark A. 2017How to knit your own Markov blanket: resisting the second law with metamorphic minds. In ***Philosophy and predictive processing: 3*** (eds Metzinger T, Wiese W). Frankfurt am Main, Germany: MIND Group.

Cleland, C. E. (2002). Methodological and Epistemic Differences between Historical Science and Experimental Science*. *Philosophy of Science*, *69*(3), 447–451.

Colombo, M., & Wright, C. (2018). First principles in the life sciences: the free-energy principle, organicism, and mechanism. *Synthese*. https://doi.org/10.1007/s11229-018-01932-w

Constant, A., Ramstead, M. J. D., Veissière, S. P. L., Campbell, J. O., & Friston, K. J. (2018). A variational approach to niche construction. *Journal of the Royal Society, Interface / the Royal Society*, *15*(141). https://doi.org/10.1098/rsif.2017.0685

Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., & Friston, K. (2020). active inference on discrete state-spaces: a synthesis. In arXiv [q-bio.NC]. arXiv. http://arxiv.org/abs/2001.07203

Dupré, J. (2020). Life as Process. *Epistemology & Philosophy of Science*, *57*(2), 96–113.

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *360*(1456), 815–836.

Friston, K. J. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301.

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nature Reviews. Neuroscience*, *11*(2), 127–138.

Friston, K. J. (2011). Embodied inference: or ``I think therefore I am, if I am what I think''. In W. Tschacher & C. Bergomi (Eds.), *The implications of embodiment: Cognition and communication* (pp. 89–125). Imprint Academic.

Friston, K. J. (2013). Life as we know it. *Journal of the Royal Society, Interface / the Royal Society*, *10*(86), 20130475.

Friston, K. J., Parr, T., & de Vries, B. (2017). The graphical brain: Belief propagation and active inference.

Network Neuroscience, *1*(4), 381–414.
Friston, K. J., & Stephan, K. E. (2007). Free-energy and the brain. *Synthese*, *159*(3), 417–458.
Friston, K. J., Thornton, C., & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology*, *3*, 130.
Hesp, C., Ramstead, M. J. D., Constant, A., & Badcock, P. (2019). A multi-scale view of the emergent complexity of life: A free-energy proposal. *Evolution & Development*. https://link.springer.com/chapter/10.1007/978-3-030-00075-2_7
Hohwy, J. (2016). The self-evidencing brain. *Noûs*, *50*(2), 259–285.
Hohwy, J. (2020). Self-supervision, normativity and the free-energy principle. Synthese. https://doi.org/10.1007/s11229-020-02622-2
Kirchhoff, M. (2015). Species of realization and the free-energy principle. *Australasian Journal of Philosophy*, *93*(4), 706–723.
Kirchhoff, M., & Froese, T. (2017). Where There is Life There is Mind: In Support of a Strong Life-Mind Continuity Thesis. *Entropy* , *19*(4), 169.
Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The Markov blankets of life: autonomy, active inference and the free-energy principle. *Journal of the Royal Society, Interface / the Royal Society*, *15*(138). https://doi.org/10.1098/rsif.2017.0792
Klein, C. (2018). What do predictive coders want? *Synthese*, *195*(6), 2541–2557.
McNamara, J. M., Green, R. F., & Olsson, O. (2006). Bayes' theorem and its applications in animal behaviour. *Oikos* , *112*(2), 243–251.
Okasha, S. (2013). The Evolution of Bayesian Updating. *Philosophy of Science*, *80*(5), 745–757.
Parr, T., & Friston, K. J. (2018). The Anatomy of Inference: Generative Models and Brain Structure. *Frontiers in Computational Neuroscience*, *12*, 90.
Parr, T., & Friston, K. J. (2019). Generalised free-energy and active inference. *Biological Cybernetics*. https://doi.org/10.1007/s00422-019-00805-w
Ramstead, M. J. D., Badcock, P. B., & Friston, K. J. (2017). Answering Schrödinger's question: A free-energy formulation. *Physics of Life Reviews*, *24*, 1–16.
Ramstead, M. J. D., Kirchhoff, M. D., & Friston, K. J. (2019). A tale of two densities: active inference is enactive inference. *Adaptive Behaviour*, 1059712319862774.
Ramstead, M. J. D., Kirchhoff, M. D., Constant, A., & Friston, K. J. (2019). Multiscale integration: beyond internalism and externalism. *Synthese*. https://doi.org/10.1007/s11229-019-02115-x
Richerson, P. J. (2018). An integrated bayesian theory of phenotypic flexibility. *Behavioural Processes*. https://doi.org/10.1016/j.beproc.2018.02.002
Sella, G., & Hirsh, A. E. (2005). The application of statistical physics to evolutionary biology. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(27), 9541–9546.
Smith, R., Friston, K., & Whyte, C. (2021, January 2). A Step-by-Step Tutorial on active inference and its Application to Empirical Data. https://doi.org/10.31234/osf.io/b4jm6
Smith, R., Kuplicki, R., Teed, A., Upshaw, V., & Khalsa, S. S. (2020). Confirmatory evidence that healthy individuals can adaptively adjust prior expectations and interoceptive precision estimates. In Cold Spring Harbor Laboratory (p. 2020.08.31.275594). https://doi.org/10.1101/2020.08.31.275594
Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., & Dessimoz, C. (2013). Approximate Bayesian computation. *PLoS Computational Biology*, *9*(1), e1002803.
Tavaré, S., Balding, D. J., Griffiths, R. C., & Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, *145*(2), 505–518.
van Es, T. (2020). Living models or life modelled? On the use of models in the free-energy principle. Adaptive Behaviour, 1059712320918678.
Wiese, W., & Metzinger, T. (2017). *Vanilla PP for Philosophers: A Primer on Predictive Processing*.

Conclusion to chapter 1

The two lessons of chapter 1 are that: (i) free-energy minimization is not sufficient for life, and (ii) that the choice of priors (and likelihood) is always at the core of Bayesian theories of cognition, including active inference and the free-energy principle. The priors that organisms are endowed with will not always guarantee that free-energy minimising behaviour will always be adaptive (in the sense of yielding positive health, survival, or reproductive outcomes). This will be the case even if the priors were selected through a process also conforming to free-energy minimization (or Bayes' Rule). This perspective on Bayesian inference leading to suboptimal, maladaptive behaviour, or simply the inability to form as a living entity, will be important to keep in mind, as this is the starting assumption of the conception of mental disorders that will interest us in computational psychiatry: mental disorders result from typical Bayesian inference that leads to maladaptive behaviour.

One thing I should stress, which might not have been clear enough from chapter 1, is that free-energy minimization is simply meant as a mathematical redescription of behavior[2]. For instance, in figure 1 of chapter 1, "expressing A" or "expressing B" just means "minimizing free energy" in a system endowed with the ability to express A or B. For that reason, it makes no sense to refer to free-energy minimization as a "condition for" life. Positing that free-energy minimization is a condition for life assumes, implicitly, counterfactual scenarios wherein one might fail to minimize free-energy and therefore fail to meet the condition for life. One does not fail to minimize free energy, any more than one fails to "conform to gravity", or that water streams fail to "conform to Navier-Stokes equations". Colin Klein in his 2018 treatment of the FEP is right when claiming that:

> *"to talk about an organisms' expectations of the world is not to propose that there are specific, concrete things which play a causal role in driving behavior. Rather, talk about minimization of free-energy and an organisms' expectations is meant to be something like a description of how whole organisms behave … the point is not to describe mechanisms but rather the overall dynamic of a system." (Klein, 2018, p.2251).*

Klein is equally right in claiming that the free-energy principle is not a satisfying explanation of life and behavior. "*Appeal to apparent tautologies should trouble you. For whatever tautologies do, they don't explain why things happen. At best, they give us reason to believe that something is the case" (Klein, 2018, 2252).* The free-energy principle gives us a good reason to believe that something is the case; that "something" being living things like us. As I argued in chapter 1, implementations of the free-energy principle under the theory of active inference – which is the modelling approach used in chapters 3 and 4 – are really what do the explanatory work. Klein is right again when saying that:

> *"it is worth keeping in mind that FEP is a starting point from which one might develop explanations, and that its defense would ultimately rest on the empirical adequacy of detailed models which spring from it. Simplicity does not count in its favor, for FEP is simple in the way that friction-free planes and infinite populations of bunnies are simple: that is, a deliberate simplification, which buys scientific fruitfulness at the cost of literal truth." (Klein, 2018, p.2553).*

[2] Some authors take FEP to be empty of empirical content (Andrew, 2022).

The free-energy principle is indeed not concerned with the literal truth. It is after truth, "within the realm of reason"[3]. The free-energy principle is not meant to tell us what is "good" and "bad" for organisms, or what they "should" be doing to remain alive; hence, Klein is – this time, almost -- right when saying that:

> *"minimizing free-energy cannot be necessary for survival either. I think this fact is often obscured by the contrast cases authors choose when they explicate FEP: the options are either being happy and healthy or else hurtling towards the bottom of the sea. But there is a large grey area between the two: life mostly requires getting by well enough, most of the time. Yet FEP places an austere set of constraints on organisms: they must minimize free energy, and so resist change, in some way that approximates optimality. We know humans aren't optimal, though. We can't be. We die." (Klein, 2018, p. 2252).*

The FEP places no constraints on organisms. Again, it is not an attempt at prescribing conditions for anything. Klein explained it beautifully himself. The FEP is just a tautology that is there to convince you that certain things are the case. Because that tautology comes with a mathematical formalism, it also tells you where you might want to start your inquiry if you want to develop mechanistic accounts of behavior that are consistent with that tautology. I understand that many philosophers have given the impression that under the FEP, organisms "*must minimize free energy, and so resist change, in some way that approximates optimality"*. But this is wrong. Rather, under the FEP, it "*must be the case that organisms minimized their free-energy if they resisted change — or rather underwent changes consistent with their life history — from t-1 to t (i.e., if we can still observe them)*. However, nothing guarantees that minimizing free-energy will lead to life and survival at t+1. Minimizing free-energy cannot be a sufficient condition for survival or for life. However, minimizing free-energy will be a necessary condition *of* life and survival. That is, it is not because one minimizes free-energy that one will survive and live. Again, if that were to be the case, one would live forever. Rather, the claim of the FEP is that whatever happens next will have been driven by free-energy minimization. Once convinced that this may be true "within the realm of reason", one can start confidently leveraging the free-energy principle to develop computational models such as those of active inference. In the next chapter, I look at how one may apply the free-energy principle to non-living systems, hoping to convince the reader that models derived from the FEP can also be meaningfully applied to modelling environmental components.

[3]This is a good time to revisit the introductory quote. I view the "truth within the realm of reason", in reference to Nietzsche, as distinct from the "literal truth", in the sense that it is not about what things are, or about how they work, but rather about one's own conviction about whether things are or work. Truth within the realm of reason obtains when uncertainty over one's credence about things is brought to a minimum. This is what happens when thinking about tautologies. Tautologies are true in the sense that we cannot conceive of them being false, even though this may mean that, nor can they ever be literally true.

Chapter 2: Extended active inference: Constructing predictive cognition beyond skulls

Introduction to chapter 2

Chapter 2 is titled “Extended active inference: Constructing predictive cognition beyond skulls”. Explicitly, chapter 2 is an attempt at providing a computational understanding of extended cognition, such as originally developed by Andy Clark and David Chalmer in "The Extended Mind" (Clark & Chalmers, 1998). To do so, we employ active inference and the free-energy principle to describe the processes whereby cognitive extensions are constructed over developmental and evolutionary time. We leverage views in cognitive niche theory, which studies the organism's niche as providing the organism with ‘instrumental intelligence’ (i.e., the ability to create and maintain cause–effect models encoded in the layout of the niche, which guide fitness influencing behaviour). In line with extended mind theory, we propose that cognitive extensions are predicted and predictable states of the niche, which support the performance of certain cognitive tasks.

From the point of view of active inference and the free-energy principle, this brings niche construction on a par with standard cognitive functions like action, perception and learning, which are all geared towards the optimization of the organism’s generative model, allowing the performance of cognitive tasks. This chapter thus argues that cognitive niche construction can be studied as a cognitive function and that cognitive niche construction can be viewed as a process of constructing, optimizing, and leveraging cognitive extensions, which are the result of ‘uploading’ some cognitive aspects of functions to the environment, to create a “coalition” between world and brain. Uploading is the process that characterizes what we have termed extended active inference. As I discuss in chapter 2, uploading is a stronger version of what is known as cognitive “off-loading” in the theory of the extended mind. The concept of “uploading”, as the term suggests, implies that the uploaded cognitive function is effectively installed in the environment, which involves creating a new function taken onboard by the environment, and that can be leveraged by the agent to perform cognitive tasks otherwise difficult if not impossible to accomplish.

Chapter 2 offers a formal and conceptual model of how the material environment comes to encode prior beliefs about organisms through cognitive uploading, and how those beliefs extend into the material environment and support cognitive functions. Crucially, the claim here is that organisms will become

dependent on these functions over evolutionary time, as they will provide survival and fitness advantages shaping future generations. In Chapter 4, the model of symptoms of depression and intervention we propose is based on this modelling strategy.

1 Introduction

This chapter reviews generic predictive approaches to niche construction to propose a specific model of cognitive niche construction under active inference. We clarify the mechanics of some important components of the cognitive niche that have yet to be addressed under active inference; namely, the functional and psychological components. We then argue for a view of extended active inference (henceforth, EAI) based on our model of cognitive niche construction. This introduction provides a definition of the key concepts we refer to in this chapter and outline of the proposed argument.

1.1 Concepts

1.1.1 The cognitive niche

In cognitive science, cognitive niche construction can be viewed as a form of instrumental intelligence whereby organisms "create and maintain cause–effect models of the world as guides for prejudging which courses of action will lead to which results" (DeVore & Tooby, 1987, p. 2010). For instance, juvenile Capuchin monkeys zero in on stones proper to nut-cracking activity by relying on traces left behind by experienced Capuchins. Residues are left on sites where successful nut-cracking activity took place, which indicates to newcomers that stones found on those sites are suitable for nut-cracking (Fragaszy, 2011). Traces, stones, and dispositions to social learning here form the ingredients of the cognitive niche as a cause-effect model.

The concept of the cognitive niche employed in cognitive science refers to the concept of the developmental, "ontogenetic niche" (West & King, 1987). The concept of the developmental niche asks a set of questions different from that of the selective niche (Stotz, 2017); it asks a question about "not what's inside the genes you inherited, but what the genes you inherited are inside of" (Stotz, 2010, p.1). This set of questions is especially interesting to study the epigenetic and behavioural sources of variations upon which selection can act. In turn, the concept of the selective niche is well suited to study the manner in which selection pressures are transformed by organisms. In evolutionary biology, the cognitive aspect of the cognitive niche refers to the effects of the developmental niche on variations that relate to cognitive functions (Stotz, 2010).

The concept of the cognitive niche we refer to here is a sort of hybrid between the concepts of the selective, developmental and cognitive niches. However, even though we rely on these parallels to make our argument, a detailed analysis of these is beyond the scope of this chapter. The set of questions that fall within our scope relate to the computational function of cognitive extensions, and the (developmental and intergenerational) process whereby this computational function emerges. For instance, from an evolutionary point of view, the

concept of the cognitive niche that interests us will focus on the evolution of cognitive extensions *per se* (in a manner akin to cumulative cultural evolution (Mesoudi & Thornton, 2018)).

The niche we consider here is made of niche construction outcomes directly relevant to an organism's activity—for example, extended phenotypes having fitness enhancing impacts (Dawkins, 1982) and "external niche inheritance" such as energetic and informational resources (Odling-Smee, 2007). External inheritance can secure the reproduction of organisms' life cycle over developmental time—for example, for beaver kits—while causing ecological cascades for other species receiving that inheritance (e.g., through modified communities). We do not include in the cognitive niche outcomes and ecoevolutionary feedback that drive evolution by either negatively impacting development (e.g., "negative" niche construction outcomes like feces), or by being "ecological cascades" that can force the exploration of the adaptive landscape (Odling-Smee, Laland, & Feldman, 2003).

The cognitive niche is sometimes studied as a psychological habitat, and sometimes as a functional habitat (cf., Bertolotti & Magnani, 2017). The psychological habitat refers to the set of organisms-niche relations that offer organisms relevant action (and perception) possibilities, also known as affordances (Gibson, 1979). The functional habitat is the set of resources that support species-specific tasks (e.g., foraging, or language and communication in humans (Clark, 2006; Whiten & Erdal, 2012)). This means that one must define the functional habitat on the background of the organism's phenotypic dispositions; for example, books are part of the functional habitats of humans because of humans' ability to read, but they are not part of the beavers' functional habitat. The psychological and functional habitats can be part of the same overall physical habitat. They simply differ in terms of their explanatory scope. The former explains psychological aspects of the organism's experience, such as perception, whereas the latter explains how the organism will rely on the niche to perform some task (e.g., foraging).

1.1.2 Active inference

Contemporary "predictive" theories of cognition include well-known theories such as predictive coding (Rao & Ballard, 1999), the Bayesian brain (Knill & Pouget, 2004), predictive processing (Clark, 2013) and the predictive mind (Hohwy, 2013), ecological enactivism (Bruineberg, Kiverstein, & Rietveld, 2016), and active inference. Active inference, in particular, is commonly used to account for cognitive phenomena such as action, decision-making, and environmental navigation (Kaplan & Friston, 2018).

Active inference assumes that an organism must entertain minimally uncertain "causal" models—that can generate effects from their causes—of the probabilistic relation between relevant types of events. Uncertainty is an information-theoretic notion that relates to Shannon information. Shannon or self-information can be quantified by measures such as surprise and entropy. Surprisal $\Im(x)$ is a measure of unlikeliness that a random variable X takes a value x, given a model m of how X was generated; that is, $\Im = -\ln P(x \mid m)$. In turn, entropy $S = E[\Im(x)]$ is the expected or weighted average of surprise over time. Crucially, the negative of surprise is also known as log model evidence or marginal likelihood $\ln P(x \mid m)$. This means that minimizing surprisal (i.e., self-information) corresponds to maximizing model evidence; which has been referred to as self-

evidencing (Hohwy, 2016). Self-evidencing over time also means minimizing uncertainty or entropy. For instance, an equal probability such as .5 and .5 of observing an outcome (e.g., $X = \{head; tail\}$ before any observation (e.g., before flipping a coin) entails a state of full uncertainty (or maximum entropy). The observation of an occurrence (e.g., after having flipped the coin) entails a full disambiguation or maximum information gain. Put another way, one defines the information gained after observing an outcome in terms of the amount of uncertainty that is resolved. Hence, a shorthand for the notion of self-evidencing is uncertainty reduction. From the standpoint of a physicist, the resolution of uncertainty corresponds to the tendency of lifelike systems to resist the second law of thermodynamics—or strictly speaking, the fluctuation theorems that apply to open systems—by placing an upper bound on their entropy or disorder.

According to active inference, to survive and reproduce when facing environmental stressors, organisms must entertain minimally uncertain models of the relation between sensory inputs they receive (e.g., "scent") and the possible environmental causes having generated these inputs (e.g., "predator"; or "mating partner"). Organisms must also model the probability of transitions among causes in the world (e.g., "predator approaching") relative to possible actions their physiology permits (e.g., "I can fly"; and "I can't swim"). In line with models of Bayes optimal foraging (Okasha, 2013), minimizing uncertainty in such causal, predictive, or generative models involves updating probabilistic mappings or Bayesian beliefs (a.k.a., learning and perceptual inference), and selectively sampling sensory inputs expected under these beliefs (a.k.a., action).

1.1.3 The extended mind

The extended mind approach to cognition (Clark & Chalmers, 1998) claims that cognitive processes can be offloaded to (i.e., reallocated to), or extended through (i.e., transformed into) components that reach beyond the system's internal states (e.g., brain states). The notion of offloading refers to the use of physical action and artefacts to manage the cognitive demand of information processing (for a review see Risko & Gilbert, 2016). Extended mind theorists suggest that the realization base of some cognitive processes (i.e., states that realize a given cognitive process) come to include reliable, accessible external states of the niche (e.g., the cellphone that functions as extended memory for recalling phone numbers (for a review see Kirchhoff & Kiverstein, 2019).

1.2 Outline

1.2.1 Current limitations

Some have drawn links between the cognitive niche construction perspective and the notion of uncertainty minimization in active inference and implicit self-evidencing. For instance, simulation studies have shown that by changing the material layout of the niche in a way that mirrors the causal models of the organism, organisms shape their sensory array in a way that is congruent with learned generative models, which entails more efficient reduction of uncertainty over development (Bruineberg, Rietveld, Parr, van Maanen, & Friston, 2018).

The mirroring, or synchronization that obtains between organisms and their niche has various feedback consequences over evolutionary time. For instance, some proposed that organisms can install in the niche cues that invite action with high epistemic value. Epistemic value relates to the ability of an action to resolve uncertainty—through the selection of actions that solicit the right sort of sensations for resolving ambiguity (e.g., looking under the streetlight or reading an instruction manual, Friston, Rigoli et al., 2015). Through external niche inheritance, salient cues with high epistemic value can be passed on as ecological legacies to guide the epistemic foraging of future generations (Constant, Bervoets, Hens, & Van de Cruys, 2018).

The process whereby organisms install epistemic cues in their environment provides a suitable mechanistic account of the notion of instrumental intelligence in cognitive niche theory. However, the mechanics of the functional and psychological dimensions of the cognitive niche remain unexplored in the literature on predictive processing approaches to cognitive niche construction (for interesting discussions of related functions see Bruineberg & Rietveld, 2014; Clark, 2013; Fabry, 2017; Ramstead, Veissière, & Kirmayer, 2016).

1.2.2 The argument

In Section 2, we unpack the functional and psychological dimensions of the cognitive niche under active inference. We argue that the cognitive niche—understood as an externally realized cause-effects model—can be modelled as a form of externally realized "shared" generative model that is leveraged and optimized by organisms to perform action related adaptive cognitive functions (e.g., decision-making, navigation, foraging). The optimization and leveraging of this shared generative model, through action and perception, is what we call extended active inference (henceforth EAI).

We argue that one can study cognitive niche construction under EAI as a bona fide cognitive function in the game of uncertainty minimization, alongside standard functions studied by active inference, such as active sensing and learning. Formally, cognitive niche construction thus construed is geared towards uncertainty minimization, thereby qualifying as a cognitive function under active inference. The functional and psychological aspects of the cognitive niche directly follow from our formalization of EAI (see Figure 2). We conclude Section 2 by presenting two case studies that illustrate the view of cognitive niche construction as a cognitive function.

In Section 3, we explain the relation between EAI, the original approach to the extended mind (Clark & Chalmers, 1998) and the diachronic approach (Kirchhoff, 2012, 2015). When viewed as a cognitive function, cognitive niche construction under active inference allows an epistemological extension of the boundaries of cognition (cf., Kirchhoff & Kiverstein, 2019). Building on Section 2, we argue that the coalition between brain(s) and world that obtains through cognitive niche construction—operate through a process of cognitive uploading (Constant, Ramstead et al., 2018). Cognitive uploading is akin to the notion of cognitive offloading in the original theory of the extended mind (Clark & Chalmers, 1998).

In contrast to the traditional notion of offloading, the notion of uploading refers to the (i) creation of novel cognitive functions (ii) that are taken on board by the cognitive niche per se; not merely managed by the niche. The uploaded cognitive functions, in contrast to off-loaded ones cannot immediately be reintegrated to skull-bound processes only. This is so because, as we will see, uploaded cognitive processes become "glued" to the organism over developmental and evolutionary time. A function is "offloaded" when individual agents re-structure their worlds so as to minimize internal processing costs and/or increase reliability. A function is uploaded when social and technological change means it is now taken care of by the niche rather than the individual. For example, most agents now store their phone numbers using smartphones rather than bio-memory. So, the whole "number storage" function (unlike the whole "remember X" function) has been assimilated into the niche. The niche into which the function has been uploaded can then be passed on to future generations for them to leverage, share and finesse that function.

The original notion of the extended mind applied, in principle, to both these kinds of cases. But the distinction is formally helpful and speaks to different webs of agent-world dynamics that evolve and alter on different spatiotemporal scales; the notion of offloading speaking to time scales spanning individual-level dynamics unfolding over real time and (neuro)developmental time scales, and the notion of uploading speaking to individual and group-level dynamics unfolding over developmental and intergenerational time scales. Uploading is a stronger species of offloading. EAI formalizes these dynamics as emergent properties of cognitive niche construction. Novel cognitive functions produced through cognitive uploading can result from gene-culture coevolutionary dynamics that "glue" organisms to those functions performed by the "trusted" niche. Uploading under EAI emphasizes the trade-off, over evolutionary and developmental time, of the deployment of on-board (neuro)biological functions for on-board (socio)environmental ones, thereby allowing metabolically efficient, though niche bound adaptive behaviour that may be favoured by selection.

Crucially, cognitive uploading endows external states of the cognitive niche with the ability to track regularities otherwise impossible to track, because they are often too complex to be learned by individual organisms. We frame affordances as uploaded proxies that track those complex causal regularities.[4] Thus, consistent with the theory of diachronic cognition (Kirchhoff, 2015), the notion of uploading can further be viewed as the process whereby agents produce cognitive extensions that gain independence from the specific individuals having produced them. Uploading differs from offloading in that the uploaded cognitive task comes to be shared by other agents. This allows the production of non-individual specific cognitive extension affording action tracking more complex regularities.

2 The functional and psychological niches under active inference

[4] Note that here we are concerned with a Gibsonian notion of affordances understood as action possibility directly perceivable in the environment. For a discussion of niche construction and pragmatic and epistemic affordances relative to mental representation of action—for example Cisek (2007) and Friston et al. (2012)—see Linson, Clark, Ramamoorthy, and Friston (2018).

Active inference explains perception and learning as processes that conform to an optimization process known as variational inference (Beal, 2003) The motivation for modelling uncertainty minimization in terms of variational inference relates to the sort of perceptual, or rather, inferential challenges faced by living systems such as humans. We have no direct access to the causes of our sensations, nor is there a one-to-one mapping between causes and sensations (Clark, 2013; Hohwy, 2016; Wiese & Metzinger, 2017); for example, a red sensation might be generated by a red traffic light, a red car, or a red jacket. These kinds of ill-posed inference problems can only be solved by appealing to prior beliefs or experience to resolve ambiguity or uncertainty; hence, the appeal to schemes such as approximate Bayesian, or variational inference.

Variational inference is a ubiquitous mathematical description of (Bayesian) belief updating that describes the formation of perceptual hypotheses that explain our sensations. Variational inference rests on a probabilistic generative model. A generative model is a probabilistic statement about a set of unobserved (hidden) variables (i.e., causes) and observed sensations (i.e., consequences), which represents an organism's predictive, or causal model of the world. A generative model is usually expressed in terms of a likelihood and a prior term:

$$\underbrace{p(s,\eta)}_{\text{generative model}} = \underbrace{p(s \mid \eta)}_{\text{likelihood}} \underbrace{p(\eta)}_{\text{prior}} \tag{1}$$

The likelihood corresponds to the probability of sensations s (e.g., "dry", or "wet") given priors about the state of the world η (e.g., "inside a burrow", or "outside a burrow"). The prior corresponds to the probability of conditions, or causes, generating the sensation (e.g., "being in or out of a burrow"), before making a sensory observation. Using variational inference, one can invert the likelihood in equation (1) to approximate the posterior probability of causes $p(\eta \mid s)$ once a sensation has been sampled. This involves the minimization of a bound on the unexpectedness of sensations (a.k.a., surprise)—called free energy—with respect to the approximate posterior, known as variational density. This density is associated with (i.e., assumed to be encoded by) internal (e.g., brain) states μ of the organism:

$$\underbrace{F}_{\substack{\text{Free}\\ \text{energy of}}} \overbrace{(s,\mu)}^{\substack{\text{sensations and}\\ \text{internal states}}} = \underbrace{D}_{\text{KLDiv.}}\left[\overbrace{q_\mu \underbrace{(\eta)}_{\substack{\text{over ext.}\\ \text{states}}}}^{\substack{\text{variational}\\ \text{density}}} \| \underbrace{p(\eta \mid s)}_{\text{true posterior}}\right] \overbrace{-\ln p(s)}^{\text{surprisal}} \tag{2}$$

In equation (2), the variational density becomes a posterior belief' about the causes of sensations (e.g., "was I in a burrow or outside a burrow η , given sensations of wetness s "). This inverse mapping—from causes to effects—corresponds to inferring the causes of sensations. In variational inference, approximating the true posterior can be described in terms of minimizing the free-energy functional $F(s,\mu)$:

$$q_\mu(\eta) = \arg\min_q F(s,\mu) \approx \underbrace{P(\eta \mid s)}_{\text{inverse mapping}} \tag{3}$$

In equation (3), this minimization has two consequences: (i) The functional becomes a tight upper bound on the unexpectedness of sensations (a.k.a., surprise); (ii) the minimization renders the variational posterior a good approximation to the true posterior. This follows because a Kullback-Leibler divergence D is always non-negative. This means, $F(s,\mu) \geq -\ln p(s)$, with equality when the divergence has been eliminated $F = -\ln p(s) \Rightarrow D = 0 \Rightarrow q_\mu(\eta) = p(\eta \mid s)$. Formally, variational inference converts an inference problem into an optimization problem as articulated by equation (3) (see Figure 1 for a summary).

Assuming that the organism's brain embodies the variational density, variational updates[5] ensure brain states encode a posterior belief about the true distribution of sensory causes and contingencies in the world, and—by the same token—the organism learns Bayes optimal priors about cause-sensation relationships. This is usually associated with experience-dependent plasticity (Friston, 2010). Hence, taken together, the dynamics described in equation (2) explain perception and learning as an optimization process, in which expectations about hidden states of the world and their relationships to each other (and sensations) are minimized with respect to free energy.

This optimization unfolds over several timescales. Neurophysiological states that underwrite inference change quickly (on a timescale of milliseconds). Neuronal connections that learn contingencies change over minutes to hours, via experience-dependent plasticity. Finally, the functional architectures that entail the generative model change over a neurodevelopmental timescale of months to years, as the phenotype becomes a sufficiently good model of its (encultured) cognitive niche (compared with the good regulator theorem (Conant & Ashby, 1970)).

Finally, in active inference, organisms are viewed as possessing priors about expected or preferred outcomes of action. This simply means that actions are selected if they bring about expected outcomes, while being geared towards minimizing expected surprise (i.e., uncertainty) about the future (Friston et al., 2014). Hence, in active inference, motor (and autonomic) functions work hand-in-hand with a perceptual inference to resolve uncertainty through the active sampling of salient, uncertainty reducing sensations, while allowing for preferred, unsurprising outcomes (green box, Figure 1).

[5] Variational updates are a ubiquitous form of Bayesian belief updating. In this paper, "beliefs" are used in the sense of belief updating and belief propagation; namely, non-propositional probability densities.

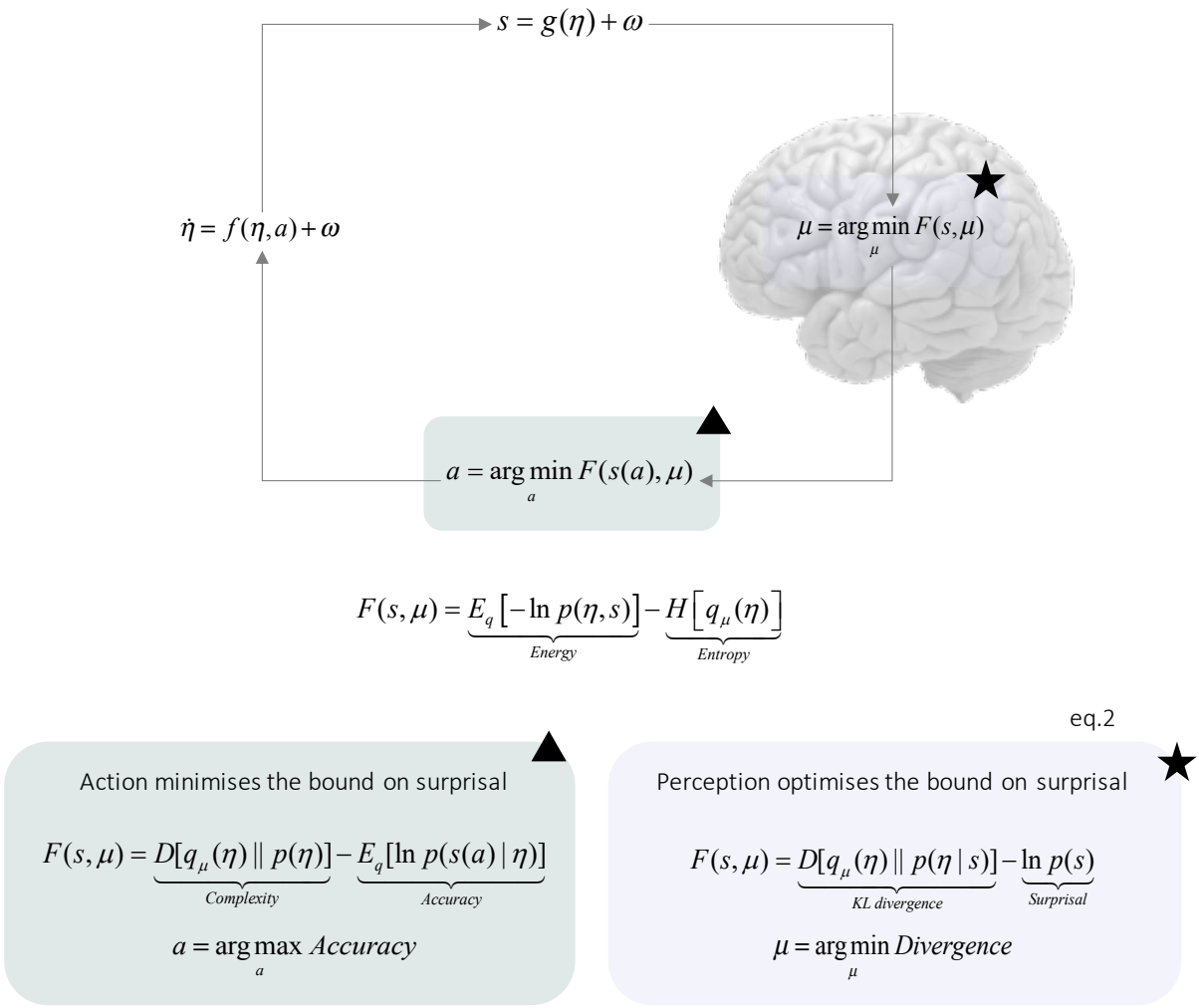


Figure 1 Action, perception, and learning under active inference.

The basic formalism corresponds to optimizing a free-energy functional of sensations and expectations encoding beliefs about hidden states of the world $F(s,\mu)$. This functional can be expressed as energy minus entropy—by analogy to free-energy in statistical physics. Various rearrangements of the free-energy functional can be used to formalize various cognitive phenomena; namely, action in the green box (triangle indicator), and perception in the purple box (star indicator). Upper panel: Sensations s and action a are the quantities that couple internal states' μ to external, hidden states in the environment η. The $\arg\min$ operator refers to variational updates—for an introduction to variational inference in relation to other inference schemes (e.g., expectation maximization) algorithms (Beal, 2003). External states are described in terms of equation of motion that include random fluctuations ω. Purple box: Perception optimizes internal states. The mathematical formulation of free-energy corresponds to equation (3) in the text. Green box: Action minimizes the free-energy bound by increasing the accuracy of sensations; for example, by selectively sampling expected sensations. Note that action does not consider posterior beliefs in the Kullback-Leibler divergence. This reflects the fact that action can only change free-energy by changing sensory inputs. When choosing among different actions, the free-energy is minimized with respect to "counterfactual" outcomes by taking the expectation of free energy, under future outcomes, given the action being evaluated. In this instance,

maximizing expected accuracy is equivalent to minimizing ambiguity. Similarly, minimizing expected complexity minimizes risk; defined as the divergence between predicted and preferred outcomes.

2.1 The cognitive niche

Changes in brain states and functional architectures optimize organisms generative (i.e., causal) model of the causal structure of their cognitive niche. Interestingly, one can use the variational formalism to model and study changes in an environment, or external states, in the same way one does for experience dependent learning in the brain (Bruineberg et al., 2018; Constant et al., 2018). We now show how this formal symmetry yields a view of cognitive niche construction as a form of environmental "learning" about the organisms hosted by the environment. On this view, organisms effectively "teach" the environment what actions they should expect (i.e., construct externally realized causal models of the effects of action—where action, from the point of view of the environment now becomes a sensory datum).

The environment is the generative process that is modelled by the generative model entailed by the phenotype. However, in virtue of the mathematical symmetry imposed by a Markov blanket (that separates internal and external states) (see Friston, 2013; Ramstead, Badcock, & Friston, 2018; Clark, 2017; Kirchhoff, Parr, et al., 2018), the environment can also be construed as a generative model of its denizens, who now becomes the processes generating outcomes for the environment. In other words, the external or environmental states play the dual role of generating outcomes for organisms, while also encoding probabilistic "beliefs" about organismal processes. We will see that one can treat the environment as inferring the cause of the "sensations" it receives from being acted upon by its denizens.

We do not claim that the formal symmetry between brain and niche dynamics entails a symmetry in construal. Rather, we employ the notion of symmetry epistemically, as a modelling "analogue" (cf., Figdor, 2018) to make sense of niche dynamics as learning dynamics under active inference. The notion of symmetry is merely an assumption that allows us to write the formal model (Figure 2) presented in this section. The added value of our model, as it pertains to this chapter, is to provide a mechanistic basis for the psychological and functional aspects of the cognitive niche. The model on offer is readily implementable in silico simulations of active inference, thereby yielding potential novel avenues for empirical research on cognitive niche construction and extended cognitive science.

Formally, what counts as a sensation in the environment are the physical actions of organisms. Then, causes of sensations can be modelled as the priors of the organism having given rise to action (i.e., niche sensations) (Ramstead, Constant, et al., 2019). Just as for the photon that hits the retina—thereby generating a sensory input leading to Hebbian learning in the brain—one can model the action of the organism encoding traces of behavioural regularities in the environment. What counts as Bayesian priors in the environment are the probability mappings between action and the organism's prior about action (Figure 2). Effectively, this closes a circle of causality; in which the niche and phenotypes are trying to learn about each other to minimize their joint free-energy or surprise. An inevitable consequence of this is that the niche and its incumbents become mutually predictable—in both directions of fit—so that the joint niche-phenotype system can be regarded as jointly self-evidencing.

Take for instance the phenomenon of desire paths. Pedestrians often leave traces in parks as they cut through the grass on their commute. Over time, these traces might become deeper, thereby telling newcomers this trail is likely to lead to outcomes preferred by the people having carved the paths; namely, people like me, who prefer or predict the same sorts of things. In so doing, desire paths encode mappings between possible actions and outcomes (e.g., "if I follow this path, I will find the café"). These mappings can have different degrees of reliability. At first, they may be ambiguous, as multiple shallow traces may encode different alternative action-outcome mappings of equal prior probability $p(\mu \mid a)$ (e.g., "this path may take me to the café"). As a path becomes more salient, it will further attract pedestrians who desire to cut through the park to reach the café, which will further consolidate the trail. Over time, assuming that people indeed find the café, the path will encode traces reducing uncertainty about the way to the café.

By analogy to perception and learning in equation (2), one can formalize cognitive niche construction as a minimization of free-energy from the point of view of the niche (see also Figure 2):

$$\underbrace{F}_{\text{Free energy of}}\ \overbrace{(a,\eta)}^{\text{organisms' actions \& states of the niche}} = \underbrace{D}_{\text{KLDiv.}}\left[\overbrace{q_\eta\ \underbrace{(\mu)}_{\substack{\text{beliefs about}\\ \text{organisms'}\\ \text{internal states}}}}^{\text{variational density}} \parallel \underbrace{p(\mu \mid a)}_{\substack{\text{true posterior of}\\ \text{organisms'}\\ \text{internal states}\\ \text{given action}}}\right] - \overbrace{\ln p(a)}^{\text{affordance}} \tag{4}$$

Equation (4) has the same form as equation (2), but with internal (sensory) and external (active) states switched around. This means that the variational density $q_\eta(\mu)$ is taken under the external states η, not internal states of the organism μ, and surprise is relative to organisms' actions. Equation (4) shows that casting changes in environmental states as self-evidencing makes the variational density—encoded by the states of the niche—a good approximation to the posterior probability over the internal states of its organisms, having observed their actions. Put another way, under this extended form of self-evidencing, the material layout of the niche will look as if it "learns" about organismal "beliefs" causing preferred action, in the same way as organisms' learn about environmental causes generating sensations.

Clearly, we are not limiting this interpretation to desire paths; in principle, any aspect of the niche can be subject to this interpretation—including cognitive, cultural and any other deontic states of the world, that is, states that tell an agent what action to select (Constant, Ramstead, Veissière, & Friston, 2019). Language itself may be considered as a kind of meta-level niche construction—a tool that allows the rapid emergence and adaptation of locally relevant niches (Lupyan and Clark, 2015)—as when someone says "the café" is under the awning across the street.

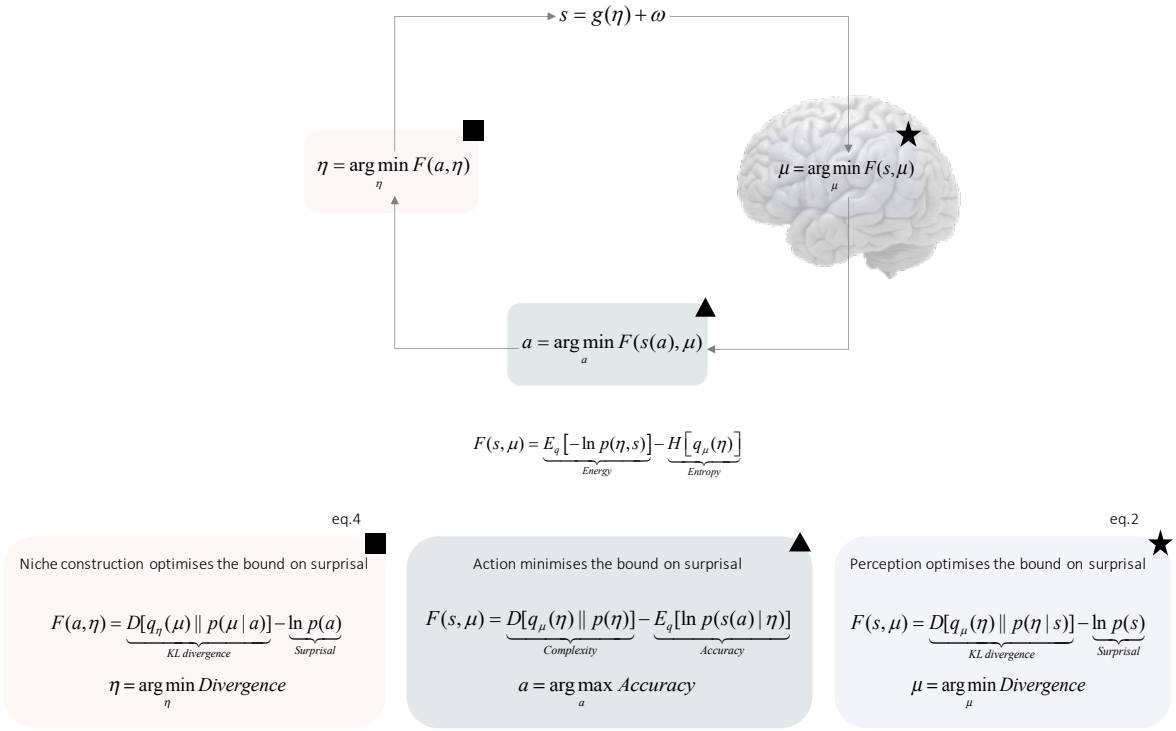


Figure 2 Cognitive niche construction and extended active inference.

As in Figure 1, internal states and action change to minimize free-energy based on sensations and internal states. Coincidentally, antisymmetric processes unfold in the niche. The key point in the figure is that all the quantities in the purple box that describe internal dynamics are inverted in the beige box—describing niche (i.e., external) dynamics. From the point of view of the niche, the action of the organism a is a "sensation", sensations of the organism s are "actions", and internal states of the organism μ are "external states". Beige box (square indicator): Cognitive niche construction as environmental "teaching" makes the environment free-energy a bound on environmental surprise. Environmental surprise here is the unexpectedness of an organism's action—or the negative log probability of encountering a particular action. This can be read as a mathematical description of affordance. In bounding surprise, the variational density of the environment ends up reflecting the most probable states of the organism, given that organism's behaviour. The expression in the beige box is reproduced in equation (4).

2.2 The psychological niche

As mentioned in the introduction, proponents of the psychological niche view the niche as a set of affordances (Rietveld & Kiverstein, 2014). In our model, the niche's free-energy bounds the surprisal of an organisms' action, and therefore can be viewed as an evidence bound on the probability of an observed action, averaging over an organisms' priors and preferences.[6] As expressed in equation (4), changes in the physical states of the niche (e.g., the production of niche construction outcomes) will optimize a bound on the surprisal of

[6] Mathematically, model evidence is also known as a marginal likelihood. This is because the evidence involves an averaging or marginalisation over the causes of some data; here, the datum is the action of an organism that is sensed by the niche.

organisms' action, which corresponds to the (negative) affordance of an action on the environment. By analogy with the creature-centric formulation of free energy, affordance just is the (log) evidence provided by an action for the niche's generative model of the active creatures it is trying to learn about.

Modelling the dynamics of a niche with the formalism in equation (4) allows us to derive a formal notion of affordances that is built into the variational formalism. Our formal interpretation supports the view according to which affordances are organism-specific action probabilities (Bruineberg & Rietveld, 2014; Tschacher & Haken, 2007) whose gradients drive niche construction, via a joint (i.e., extended) minimization of variational free energies. Importantly, our model clarifies the manner in which the concept of affordance may be implemented in *in silico* simulation studies and empirical research under active inference, as it makes this notion readily implementable with the freely available simulation routines employed in active inference research (see the various DEMOs of the Statistical Parametric Mapping 12, MATLAB toolbox at, fil.ion.ucl.ac.uk/spm/software/spm12/). Artificial data acquired from in silico simulations of affordance production and leveraging could then be compared with empirical data (cf., Mirza et al., 2019; Cullen et al., 2018) to test hypotheses about EAI as an emergent property of cognitive niche construction under active inference (e.g., in a foraging or navigation task).

The notion of extended active inference or self-evidencing reflects the extensive aspect of free energy; namely, the free-energy of two systems (i.e., organism and niche) is just the sum of their respective free energies, conditioned upon the (i.e., sensory and active) states they share (Bruineberg et al., 2018). The psychological niche can thus be viewed as a state space of invitations to act, with peaks and valleys that correspond to the most and least probable (and thereby adaptive) actions given the priors and phenotypic preferences of organisms "like me" having constructed the niche in first place.

2.3 The functional niche

Active inference assumes that cognitive functions are in the game of optimizing an organism's generative model about the cause of its sensations. This amounts to minimizing free-energy or maximizing model evidence through variational updates (i.e., perception—purple box Figure 1), and to the selective sampling of expected sensory information (i.e., action—green box Figure 1). We now argue that cognitive niche construction (beige—box Figure 2) can be framed as a cognitive process, as construed by active inference, that optimizes an organism's generative model vicariously as part of an extended process of self-organization or self-evidencing. Niche and organisms can be meaningfully studied as trying to optimize their respective models of each other.[7]

[7] It might be argued that as this process unfolds, brains really do (due to their telos) alter so as to fit the world but that it merely appears as if the world alters so as to fit the brain. If I press my punch into the wax, it may seem odd to depict the wax as actively modelling my punch. However, if I consider the wax in relation to my hand, my hand in relation to the letter, the letter in relation to the mailman, the mailman in relation to the postal service, and the postal service in relation to my friend to whom my sealed letter is destined—all of which, just like the wax, are external states to my brain, attributing to that entire ecology the ability to engage in something like active modelling—as well as a deep hierarchical structure—starts to become more tenable (Ramstead et al., 2019). The point here is that neither of the internal or external (sub)components of the brain-wax system exist in isolation. "Oddness" arises when considering the wax as isolated from its embedding, just like oddness would arise from considering the motility of a single dendrite in isolation from the rest of its neural ecology. We do not have more space to elaborate on this argument. All that matters for our current purposes is the availability of an essentially

The take-home message of Section 2 is illustrated in Figure 2; namely, one can study the niche as the organism's generative process, or a generative model of the organism—in the sense of DeVore and Tooby (1987, p. 2010)—that implicitly learns about organismal priors and preferred behaviour. This explains why resources encoded by acting on the functional niche come to cue or afford adaptive action. As argued above, resources in the cognitive niche cue actions that were selected by conspecifics in the past. Once learned, cues—conveyed as affordances—gear the organism towards selecting actions that will tend to be adaptive (more often than not), relative to the task that entailed the carving of the niche in first place. Task specific, adaptive actions thus just are actions that bring about sensory information that are expected under the sort of priors and preferences that constitute the phenotype of organisms "like me" (Constant, Ramstead, et al., 2018; Friston, 2010).

2.4 Case study

In this subsection we unpack the view of cognitive niche construction as a cognitive function through a well-known case study in niche construction theory: The phylogeny of freshwater kidneys in common earthworms (*Lumbricus terrestris*). We take this case study as an illustration of the way earthworms optimize their generative model by encoding reliable cause-effect relationships in their environment. We then provide some examples of the effect of cognitive niche construction as a cognitive function in humans by focusing on a discussion of spicing in food preparation.

Common earthworms are phylogenetically related to aquatic freshwater worms. Freshwater worms have kidneys that remove excess water from their body. This trait is consistent with aquatic environmental conditions but far from being adaptive for terrestrial life conditions, as water is limited, and water conservation should be the norm. Thus, all things being equal—in the world of natural selection—common earthworms should have evolved water-balance organs that favour water conservation. However, common earthworms still have roughly the same freshwater kidneys as their ancestors. A plausible explanation for this is that the niche construction undertaken by earthworms might have tipped the balance in evolution. By constructing—and inheriting—semi-aquatic environments like moist soils, common earthworms might have softened selection pressures on water-balance organs (Satchell, 1983; Scott Turner, 2009). Put another way, the niche became part of common earthworms' solution space to the challenge of having water removing organs in dry environments. The niche then allowed economies of "evolutionary money" to be spent on biological adaptations (e.g., selecting for water conserving organs); thereby explaining, in part, the evolutionary trajectory having led to the current phenotype.

In the parlance of active inference, the niche of common earthworms functions to inform a predictive (or generative) model of the relation between states of the world (e.g., "in a burrow", or "outside a burrow") and sensory outcomes (e.g., "wetness", or "dryness"), cueing earthworms about relevant cause-effect relationships (Christopoulos & Tobler, 2016). The networks of burrows that generations of earthworms

symmetric formalism within which to model processes of mutual modelling between agents and their niche, which reflects genuine, relevant, and perhaps more easily conceivable forms of mutual adaptation.

constructed (and inherited) came to afford adaptive action in the sense that engaging them most likely led to locations affording a priori preferred level of wetness. In other words, cognitive niche construction outsourced the computation of adaptive action to the environment *per se*. Calling on recent numerical analyses and theoretical treatments of active inference in decision-making, we speculate that a consequence of this is that earthworms could simply rely on the action afforded by the niche to avoid computing action that would fulfil their evolutionary (prior) preferences for wet soil, which would soften selection on water balance organs.

Cognitive niche construction here operates through (i) the increase in performance enabled by the outsourcing of the computation to the niche and (ii) the absence of an adaptation due to niche construction. First, constructing cognitive niches so as to make them more predictable (i.e., navigable) enables the organism to reduce model complexity[8] by constraining the variety of sensory causes that the organism has to entertain (Sengupta, Stemmler, & Friston, 2013). This allows the enhancement of performance for exploitative, fitness related behaviour (Friston et al., 2016). Indeed, tracking the potential causes of sensations in a constantly fluctuating world is costly as it requires to entertain multiple counterfactual priors (e.g., "will I end up in a wet environment if I move left, right, up and, down, etc.?"). Outsourcing the computation of these counterfactuals to the niche can be expected to increase performance in terms of both thermodynamic and inferential efficiency. Second, the enhancement of performance may be reflected in more efficient reaction times during exploitative behaviour, which would favour the reproduction of a phenotypes that call on the predictability afforded by the niche.[9]

In earthworms, the circular causality over developmental and evolutionary time scales between the optimization of generative models through environmental modifications and the coupling to those environmental modifications over evolutionary time—may explain the softening of selection on things like water absorbing organs. This may be viewed as a form of developmental constraint on selection; that is, the strategy of outsourcing the computation became locked-in, because of the advantaged it provided, yet, to the cost of a phenotype that would heavily rely on this strategy (e.g., a phenotype that would not possess the right kidney). The phylogenetic trajectory of earthworms exemplifies the phenomenon of cognitive uploading discussed in the introduction of this chapter. Uploading here, operates through the saving on metabolic resources through the reliance on epistemic cognitive extensions that take on-board functions such as planning, which is typically internally realized. Over multiple generations, this comes at the cost of becoming "evolutionarily glued" to those cognitive extension. Put bluntly, cognitive niche construction smartens the world of the earthworm, so that its physiology can remain dumb yet optimal in peace (Clark, 1998).

The example of the earthworms speaks to the fact that characteristic behavioural patterns or components of phenotypes (extended or else) will emerge from the construction of the cognitive niche and its impacts on evolution and development. Cognitive uploading could also allow one to formalize the computational architecture of the human phenotype. For instance, the inheritance of epistemic resources over evolutionary

[8] Complexity here, is used in the technical sense of statistical complexity or complexity cost. Model evidence (i.e., negative free energy), is expressed as accuracy minus complexity. This means that self-evidencing is necessarily optimized when accurate model predictions are maintained with minimum complexity (see equation (1)).
[9] Technically, this is expressed in terms of a variational principle of least action. In other words, the imperative for self-evidencing is to minimize the time average of free energy, where this time average is known in physics as an action (not to be confused with the action associated with acting on the niche).

time and the re-enactment of the practices invited by these resources over development underwrites the phenomenon of tradition; understood as learned a new behaviour supported by socio-cultural practices (Fragaszy & Perry, 2003). In humans, traditions and associated artefacts undergo processes of cultural evolution (Boyd & Richerson, 1988), which enable intergenerational groups to converge on adaptive repertoires of tools, technologies, rituals, and so forth, that have been filtered by generations of conspecifics (for a review, see Laland, 2018).

Evolved traditions enable the success of complex cognitive tasks, while leaving the structure of the causal models — to which the success of these tasks relate — unbeknownst to the agent (Fragaszy, 2011). In his book, *The secret of our success* (Henrich, 2015), Harvard anthropologist Joseph Henrich provides a series of such simple examples in which traditions track cause-sensation relationships, otherwise impossible to track; thereby securing adaptive low-cost behaviour. One such example is the use of spices in food processing. Spices generally have no nutritional value and are often made of aversive active ingredients. Yet, many humans use them abundantly because some of those active agents turn out to kill foodborne pathogens present, for instance, in widely consumed food like meat; something that is generally unknown to people having acquired and reproducing the practice, yet that is highly beneficial to them. Traditions of spicing per se come to model hidden causes whose structure could not be discovered by individuals alone over their lifespan. In the spirit of Henrich's reflection, culture makes us smart.

From the point of view of cognitive niche construction as a cognitive function under active inference, spicing traditions are intergenerational group-level strategies to track the complex multidimensional causal relationship between spices, active agents, foodborne pathogens, and meat consumption behaviour, which supports the reproduction of the behavioural phenotype. Spicing traditions thus can be viewed as encoding a generative process constructed by multiple generations about what compound is deleterious to what pathogen, and what pathogen is deleterious to humans, and what spices should be consumed. Enculturated agents, then, become coupled to this generative process which secures adaptive food processing.

Crucially, it is the generative process embodied by the tradition per se that tracks this complex causal relationship, not individual agents. In responding to affordances (a.k.a., epistemic cues of least improbable action engaged by conspecifics; cf., Figure 2) such as those offered by artefacts of traditions, organisms like us manage to succeed implicitly in tasks for which causal models are too complex and too costly to be taken on-board. Tradition endows individuals with the ability to read into deep hidden causal regularities. In a scaffolded fashion (cf., Sterelny, 2010), the structure of extended cognition is explained formally in terms of intergenerational learning dynamics in the generative process produced by generations of niche constructing agents (i.e., people participating and reproducing the tradition), and by the enculturation of individuals' generative models through the learning of the epistemic cues (a.k.a., affordances) in the generative process.

3 Extended active inference

Over developmental time, smartening the world through cognitive niche construction operates through processes akin to that of cognitive offloading, such as studied by the extended approach to cognition. From

the perspective of active inference, cognitive niche construction brings the notion of offloading a step further. As we have seen with the earthworm and food preparation examples, cognitive uploading through cognitive niche construction entails outsourcing the inference over future outcomes to epistemic cues of the niche (a.k.a., affordances). Thus, through niche construction, organisms manage to upload self-evidencing processes directly to the structure of the generative process.

Uploading entails more than relying on physical action and artefacts to support, or help carry out, cognitive functions. The evaluation of expected surprise drives action selection. Self-evidencing refers to the process of minimizing the bound on surprisal (a.k.a., negative log model evidence) through perception (optimizing the bound) and action (minimizing the bound) (cf., Figure 1); hence cognitive uploading through cognitive niche construction outsources part of the computation of self-evidencing processes (those relating to action). Put simply, cognitive uploading helps agents to minimize the bound on surprise.

In the remainder of this chapter, we explain the manner in which the above formalism grounds EAI and generalizes two varieties of claims on extended cognition; the original approach to the extended mind (Clark & Chalmers, 1998) and its recent reinterpretation as diachronic cognition (Kirchhoff & Kiverstein, 2019). We show how EAI supports the theory of the extended mind by providing mechanistic explanation of well-known concepts such as the parity principle, functional isomorphism, epistemic action, and diachronic cognition. We do not engage the many debates surrounding the varieties of extended cognition. This is well beyond the scope of this chapter. Rather, the hope is to provide future researchers with a formal apparatus to make progress in these debates by showing how the varieties of claims on extended cognition may be formally expressed in EAI; a lingua franca of sort such as summarized in Figure 2.

3.1 The extended mind under EAI

3.1.1 Parity principle under EAI

The original theory of the extended mind decomposes into three features. The first is a parity principle. The role of the parity principle in the theory of the extended mind is to first help us to conceive of the view of the mind as being extended into external vehicles; the parity principle is "a mean of freeing ourselves from mere bio-chauvinistic prejudices" (Clark, 2005, p. 2). The parity principle states that:

> *If … a part of the world functions as a process which, were it done in the head, we would have no hesitation in recognizing it as part of the cognitive process, then that part of the world is … part of the cognitive process.* (Clark & Chalmers, 1998, p. 8)

If we agree that the function performed by an external state during a cognitive task would qualify as a bona fide cognitive function "were it done in the head", then that external state in question ought to be considered as potentially an integrative part of the cognitive architecture of the cognitive system. This principle is vindicated by the formalism of EAI presented in this chapter; as we have shown, the description of the dynamics underlying learning in the generative process are formally equivalent to the learning in the

generative model. Of course, one must consider the part of the generative process that is coupled to the generative model through cognitive niche construction.

3.1.2 Functional isomorphism under EAI

The parity principle entails the second feature of the theory of the extended mind, which is the notion of a potential functional isomorphism between some internal and some external states (Sutton, 2010). Functional isomorphism stresses that internal and external states have to be seen as equivalent with regard to the basic properties of cognition. For instance, under certain conditions, a notebook might very well play the same coarse-grained functional role, or epistemic function than biological memory implemented by patterns of neuronal connections in the brain. When looking for coarse-grained parity between internal and external cognitive resources, it has been suggested that external resources should meet the requirements of "glue and trust" so that the resource is available when needed (like bio-memory) and not subject to constant agentive scrutiny—to ensure it is working as it should (again, like bio-memory).

From the point of view of EAI, the trust condition is guaranteed by the uploading process whereby the agent learns to engage epistemic cues of the generative process. This entails trading-off on-board neurocognitive functions for on-board environmental ones. The benefit is the increased performance, though at the cost of increased dependence on the environment. The glue condition is guaranteed by the increased performance that underlies the uploading. For instance, the earthworm is "glued" to its inheritance of burrows and moist soil because of the constraints burrows and soils have operated on earthworm's phylogeny. We can imagine how an individual would become "glued" to her environment in a similar fashion, though over developmental time scale. For instance, we can imagine an individual that would carve out a path on her commute, and over time, come to heavily rely on that path to arrive to the office on time. The short cut may free up her schedule enough for her to get used to stop at the café to grab a quick espresso during her commute. Then, the individual might stop buying coffee for her kitchen; this would surely simplify the planning of her weekly stop at the grocery store anyway. This, however, would come at the cost of sticking to her path and the espresso it affords. In this hypothetical scenario, the trade-off that glues the individual to her environment is instantiated by the acquisition of a habit whose robustness relies on (un)learned states of the generative model and learned states of the generative process.

3.1.3 Epistemic action under EAI

The original theory of the extended mind argues that the environment on which cognitive agents rely enables them to perform epistemic actions (Clark & Chalmers, 1998). Epistemic actions are defined as actions that ease or optimize cognitive tasks by reducing the memory load required to perform a task (space complexity); by simplifying the computational processing procedure (time complexity); by minimizing the probability of error outcomes (success probability) (Kirsh & Maglio, 1994). A notebook, for instance, can be viewed as supporting, and easing the task of, say, making it to your multiple appointments throughout the week, as it will encode relevant information like addresses (i.e., save on space complexity), provide a structure like a schedule for

knowing when your appointments are, and how best to coordinate them (minimizes time complexity), and will probably increase your chances of making it on time (increase success probability). These intuitions are formalized by the process of uploading from the point of view of EAI, but in addition, by accounting for the relation between all these advantages. Space complexity corresponds to reduced numbers of counterfactual scenarios that one has to model, which naturally entails minimizing the probability of errors (i.e., the more complex the generative model is, the more likely it is to overfit), and by the same token increase performance in terms of time complexity of computation.

Another (complementary) way to view the picture of extended minds under EAI is to note that neutrally supported estimations of salience (a.k.a., expected surprise) help select actions that can purposefully roll in cognitive operations flowing through bio-external resources. That rolling in can be internally instigated (e.g., as when I retrieve my smartphone to ensure I do not miss my flight). My purposeful rolling in can also be cued by the external resources themselves (e.g., if I set an alarm for two hours before the flight). In that case, the drive or readiness to act to minimize my uncertainty (or to increase the precision of my beliefs about the time) will reduce, as my expectation about future surprise, or salience will decrease (e.g., "I will not feel the urge to keep verifying the time at short intervals because I will know when to access my phone"). Here, salience is managed by the cell phone, as trustworthy information is made reliably available. Crucially, the internal flux of precision (i.e., uncertainty in my beliefs) is resolved by the externally structured flow of epistemic (i.e., salience minimizing) action that serves to improve the long-term fit between my actions and my goals, as well as the cost of computing these long-term goals. Temporary coalitions of internal and external resources are thus recruited in the same way as are temporary purely inner coalitions, which likewise emerge as varying patterns of effective inner connectivity controlled by fluctuating precision and salience estimations (see Clark, 2016, Chapters 8 and 9).

As we will see below, both the long-term built environment and the cultural milieu further scaffold this process, nesting our individually extended minds inside larger co-constructed niches that likewise extract, flag, and cue optimal (i.e., expected free-energy minimal) action.

3.1.4 Diachronic cognition under EAI

The diachronic perspective casts cognitive systems as extended, not only in terms of their spatial realization, beyond the spatial scales at which the agent exists but also in terms of its temporal realization, to (legacy) scales that cognition occupies historically, and in the context of cultural practices in the here-and-now. Cognitive assemblies are formed and maintained diachronically, beyond the local organism-centered boundaries of individuals (Kirchhoff, 2012, 2015, 2018; Malafouris, 2015; Stotz, 2010). Cognitive assemblies are decentralized systems, or networks of human-and-nonhuman agencies (Latour, 1993), whose constitutive causal relationship depends upon self-organized processes distributed across the network they constitute (cf., Figure 2 for a simple environment-organism system).

The standard example used to explain diachronic cognition is that of the Elizabethan theatre companies (Tribble, 2005). Tribble explains how players of the Elizabethan theatre companies during the 16th century

would manage to perform multiple different plays per week without being able to rehearse due to time limitations. The ability of the actors to memorize how to perform plays depended on patterned sociocultural practices mediated by material artefacts populating the stage (e.g., stage doors, playing platform, plots, and scripts), and a cross-generational apprenticeship system (Sutton, 2010) allowing the (re)acquisition of the skills necessary to leverage the informational structure afforded by the augmented stage.

Under EAI, this allows the environment to learn shared preferences and narratives under the form of epistemic cues but only to the extent they are preserved by organisms acting on that environment. Each member of the theatre company engages the diachronic assembly as a generative process from the stance of their generative model. For each individual, other people and artefacts come to encode affordances that indicate what action will be successful because of the ongoing uploading of epistemic cues to the generative process through the apprenticeship practice. As in the earthworm case study discussed above, learning how to leverage these cues allows each individual to limit the complexity of their generative model, thereby enhancing performance (e.g., memory recall, reaction times, etc.) and allowing patterned, low-cost action selection.

4 Concluding remarks

The model of cognitive niche construction proposed in this chapter offers a formal apparatus for the study of non-brain-based factors in cognition. This chapter argued that cognitive niche construction could be viewed as a bona fide cognitive function. Then, we sketched some examples of how this model could be used to give a formal grip to theories of the extended mind and diachronic cognition.

The point stressed in this chapter was that cognitive niche construction can be studied as a shared cognitive function enabling organisms to track—often implicitly and at low cost—cause–effect relationships otherwise difficult, if not impossible to track; notably, relationships wherein the hidden causal structure is highly volatile, or wherein the hidden causal structure is too complex to be learned solely based on sensations available to the biological sensory apparatus of a single phenotype. From the point of view of extended active inference, all cognitive functions are in the game of tracking causal regularities, and there is no principled reason to restrict this process to the boundaries of skin, skull, or even individual agents.

References

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference* [Doctoral dissertation, Ph.D., University of London]. https://cse.buffalo.edu/faculty/mbeal/thesis/beal03.pdf

Bertolotti, T., & Magnani, L. (2017). Theoretical considerations on cognitive niche construction. *Synthese*, *194*(12), 4757–4779.

Boyd, R., & Richerson, P. J. (1988). *Culture and the Evolutionary Process*. University of Chicago Press.

Bruineberg, J., Kiverstein, J., & Rietveld, E. (2016). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, 1–28.

Bruineberg, J., & Rietveld, E. (2014). Self-organization, free-energy minimization, and optimal grip on a field of affordances. *Frontiers in Human Neuroscience*, *8*, 599.

Christopoulos, G. I., & Tobler, P. N. (2016). Culture as a response to uncertainty: Foundations of computational cultural neuroscience. In J. Y. Chiao, S.-C. Li, R. Seligman, & R. Turner (Ed.), *The Oxford handbook of cultural neuroscience* (pp. 81–104). Oxford University Press.

Cisek, P. (2007). Cortical mechanisms of action selection: the affordance competition hypothesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *362*(1485), 1585–1599.

Clark, A. (1998). *Being There: Putting Brain, Body, and World Together Again*. MIT Press.

Clark, A. (2005). Intrinsic Content, Active Memory and the Extended Mind. *Analysis*, *65*(1), 1–11.

Clark, A. (2006). Language, embodiment, and the cognitive niche. *Trends in Cognitive Sciences*, *10*(8), 370–374.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioural and Brain Sciences*, *36*(03), 181–204.

Clark, A. (2015). *Surfing uncertainty: prediction, action, and the embodied mind*. Oxford University Press.

Clark, A. (2017). How to Knit Your Own Markov Blanket: In T. K. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing*. MIND Group.

Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis*, *58*(1), 7–19.

Conant, R. C., & Ashby, R. W. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, *1*(2), 89–97.

Constant, A., Bervoets, J., Hens, K., & Van de Cruys, S. (2020). Precise Worlds for Certain Minds: An Ecological Perspective on the Relational Self in Autism. *Topoi. An International Review of Philosophy*, *39*(3), 611–622.

Constant, A., Ramstead, M. J. D., Veissière, S., & Friston, K. J. (2019). Regimes of Expectations: An active inference Model of Social Conformity and Decision Making. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2019.00679

Constant, A., Ramstead, M. J. D., Veissière, S. P. L., Campbell, J. O., & Friston, K. J. (2018). A variational approach to niche construction. *Journal of the Royal Society, Interface / the Royal Society*, *15*(141). https://doi.org/10.1098/rsif.2017.0685

Cullen, M., Davey, B., Friston, K. J., & Moran, R. J. (2018). active inference in OpenAI Gym: A Paradigm for Computational Investigations Into Psychiatric Illness. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging*, *3*(9), 809–818.

Dawkins, R. (1982). *The extended phenotype*. Oxford University Press.

Fabry, R. E. (2017). Predictive Processing and Cognitive Development. In T. K. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing*. MIND Group.

Figdor, C. (2018). *Pieces of Mind: The Proper Domain of Psychological Predicates*. Oxford University Press.

Fragaszy, D. (2011). Community Resources for Learning: How Capuchin Monkeys Construct Technical Traditions. *Biological Theory*, *6*(3), 231–240.

Fragaszy, D. M., & Perry, S. (2003). Towards a biology of traditions. In *The Biology of Traditions: Models and Evidence* (pp. 1–32). Cambridge University Press.

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nature Reviews. Neuroscience*, *11*(2), 127–138.

Friston, K. J. (2013). Life as we know it. *Journal of the Royal Society, Interface / the Royal Society*, *10*(86), 20130475.

Friston, K. J., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). active inference and epistemic value. *Cognitive Neuroscience*, *6*(4), 187–214.

Friston, K. J., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2014). The anatomy of choice: dopamine and decision-making. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *369*(1655). https://doi.org/10.1098/rstb.2013.0481

Friston, K. J., Shiner, T., FitzGerald, T., Galea, J. M., Adams, R., Brown, H., Dolan, R. J., Moran, R., Stephan, K. E., & Bestmann, S. (2012). Dopamine, affordance and active inference. *PLoS Computational Biology*, *8*(1), e1002327.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin.

Henrich, J. (2015). *The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.
Hohwy, J. (2013). *The predictive mind*. Oxford University Press.
Hohwy, J. (2016). The self-evidencing brain. *Noûs*, *50*(2), 259–285.
Kirchhoff, M. (2012). Extended cognition and fixed properties: steps to a third-wave version of extended cognition. *Phenomenology and the Cognitive Sciences*, *11*(2), 287–308.
Kirchhoff, M. (2015). Extended Cognition & the Causal-Constitutive Fallacy: In Search for a Diachronic and Dynamical Conception of Constitution. *Philosophy and Phenomenological Research*, *90*(2), 320–360.
Kirchhoff, M. D., & Kiverstein, J. (2019). *Extended Consciousness and Predictive Processing: A Third-Wave View*. London, UK: Routledge.
Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The Markov blankets of life: autonomy, active inference and the free-energy principle. *Journal of the Royal Society, Interface / the Royal Society*, *15*(138). https://doi.org/10.1098/rsif.2017.0792
Kirsh, D., & Maglio, P. (1994). On Distinguishing Epistemic from Pragmatic Action. *Cognitive Science*, *18*(4), 513–549.
Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719.
Latour, B. (1993). *Petites leçons de sociologie des sciences*. Découverte.
Linson, A., Clark, A., Ramamoorthy, S., & Friston, K. J. (2018). The active inference Approach to Ecological Perception: General Information Dynamics for Natural and Artificial Embodied Cognition. *Frontiers in Robotics and AI*, *5*, 611.
Lupyan, G., & Clark, A. (2015). Words and the World: Predictive Coding and the Language-Perception-Cognition Interface. *Current Directions in Psychological Science*, *24*(4), 279–284.
Malafouris, L. (2015). Metaplasticity and the Primacy of Material Engagement. *Time and Mind*, *8*(4), 351–371.
Mesoudi, A., & Thornton, A. (2018). What is cumulative cultural evolution? *Proceedings of the Royal Society B: Biological Sciences*, *285*(1880), 20180712.
Mirza, M. B., Adams, R. A., Friston, K., & Parr, T. (2019). Introducing a Bayesian model of selective attention based on active inference. *Scientific Reports*, *9*(1), 13915.
Odling-Smee, J. (2007). Niche Inheritance: A Possible Basis for Classifying Multiple Inheritance Systems in Evolution. *Biological Theory*, *2*(3), 276–289.
Odling-Smee, J., Laland, K. N., & Feldman, M. W. (2003). *Niche Construction: The Neglected Process in Evolution*. Princeton University Press.
Okasha, S. (2013). The Evolution of Bayesian Updating. *Philosophy of Science*, *80*(5), 745–757.
Ramstead, M. J. D., Badcock, P. B., & Friston, K. J. (2018). Answering Schrödinger's question: A free-energy formulation. *Physics of Life Reviews*, *24*, 1–16.
Ramstead, M. J. D., Constant, A., Badcock, P. B., & Friston, K. J. (2019). Variational ecology and the physics of sentient systems. *Physics of Life Reviews*, *31*, 188–205.
Ramstead, M. J. D., Kirchhoff, M. D., Constant, A., & Friston, K. J. (2019). Multiscale integration: beyond internalism and externalism. *Synthese*. https://doi.org/10.1007/s11229-019-02115-x
Ramstead, M. J. D., Veissière, S. P. L., & Kirmayer, L. J. (2016). Cultural affordances: scaffolding local worlds through shared intentionality and regimes of attention. *Frontiers in Psychology*, *7*, 1090.
Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.
Rietveld, E., & Kiverstein, J. (2014). A rich landscape of affordances. *Ecological Psychology: A Publication of the International Society for Ecological Psychology*, *26*(4), 325–352.
Risko, E. F., & Gilbert, S. J. (2016). Cognitive Offloading. *Trends in Cognitive Sciences*, *20*(9), 676–688.
Satchell, J. E. (1983). *Earthworm Ecology: From Darwin to Vermiculture*. Chapman and Hall.
Scott Turner, J. (2009). *The Extended Organism: The Physiology of Animal-Built Structures*. Harvard University Press.

Sengupta, B., Stemmler, M. B., & Friston, K. J. (2013). Information and efficiency in the nervous system--a synthesis. *PLoS Computational Biology*, *9*(7), e1003157.
Sterelny, K. (2010). Minds: extended or scaffolded? *Phenomenology and the Cognitive Sciences*, *9*(4), 465–481.
Stotz, K. (2010). Human nature and cognitive–developmental niche construction. *Phenomenology and the Cognitive Sciences*, *9*(4), 483–501.
Stotz, K. (2017). Why developmental niche construction is not selective niche construction: and why it matters. *Interface Focus*, *7*(5), 20160157.
Sutton, J. (2010). Exograms and Interdisciplinarity: history, the extended mind, and the civilizing process. In R. Menary (Ed.), *The Extended Mind* (pp. 189–225). MIT Press.
Tooby, J., & DeVore, I. (1987). The Reconstruction of Hominid Behavioural Evolution Through Strategic Modeling. In W. G. Kinzey (Ed.), *The Evolution of Human Behaviour: Primate Models* (pp. 183–237). SUNY Press.
Tribble, E. B. (2005). Distributing Cognition in the Globe. *Shakespeare Quarterly*, *56*(2), 135–155.
Tschacher, W., & Haken, H. (2007). Intentionality in non-equilibrium systems? The functional aspects of self-organized pattern formation. *New Ideas in Psychology*, *25*(1), 1–15.
West, M. J., & King, A. P. (1987). Settling nature and nurture into an ontogenetic niche. *Developmental Psychobiology*, *20*(5), 549–562.
Whiten, A., & Erdal, D. (2012). The human socio-cognitive niche and its evolutionary origins. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1599), 2119–2129.
Wiese, W., & Metzinger, T. K. (2017). Vanilla PP for Philosophers: A Primer on Predictive Processing. In T. K. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing*. MIND Group.

Conclusion to chapter 2

The goal of chapter 2 was to show formally that one can study the way in which the material world changes over time — its behaviour — in the same way that one can study how nervous systems change under the free-energy principle. In the same way that living organisms perceive, act and learn by minimizing free energy, nonliving things (consistently acted upon) can 'perceive', 'act' and 'learn' by having their free-energy minimized through organisms' actions. Of course, the meaning of perception, action and learning here denote their formal referent under the free-energy principle. I am aware that this is a strong claim; especially if you (still) think that minimizing free-energy is a sufficient condition for life, as often seems to be the case. If minimizing free-energy is a sufficient condition for life, and the material environment minimizes free energy, should we consider the material environment as being alive? No, we shouldn't, or at least, this is not a question that we can settle with the free-energy principle. This is the problem chapter 1 was designed to avoid.

Again, I am not advocating for panpsychism, nor am I making any ontological claim. And in the same way, when I talk about perception, action and learning of the environment, I do not use those terms as denoting a first-person, subjective experience. Rather, these denote very general computational processes that can be described by free-energy minimization. Depending on their scale (e.g., in the biosphere of fig. 1 in the introduction), perception, action and learning will be attributes of systems that are more or less complex, with different material implementations, which we can imagine will account for the richness of those processes at those different scales. Again, the ultimate goal of this dissertation is to come up with a model of mental disorders that allows one to take seriously the relation between all the levels of the biosphere (cf. fig. 1, introduction). To achieve that, one needs a vocabulary to talk about the processes underwriting the behaviour

of the systems at each level. free-energy minimization under active inference may be viewed as such a universal vocabulary.

While talk of perception, action and learning by the environment is simply a way to denote processes that may have a counterpart across the biosphere, there is a genuine distinction between those systems that I failed to emphasize in chapter 2. This distinction was raised multiple times by my colleagues and mentors, notably by Laurence Kirmayer, and only recently did I manage to understand its role in the model presented in chapter 2. That distinction, as presented by Prof. Kirmayer, is that living agents will asymmetrically act on the environment relative to the way the environment will act on them. The symmetry and asymmetry are with respect to the relation of living agents with their environment and the relation of an environment with the living agents it hosts. Individual agents can modify the environment according to some goal, plan or intentions, whereas the environment cannot modify the agent that way. That is the asymmetry. Put more simply, an agent has the kind of autonomy that is realised by the capacity to plan, whereas the environment does not plan in the sense of active inference. I think that this is correct. However, the story requires more nuance.

At the level of the group, actions can lack goal orientedness while appearing organized. For instance, nobody "plans" to carve out a desire path in urban public parks. When considering a group of agents embedded in an environment, one can recover symmetry (i.e., the idea that both directions in the relation lack intentionality), as arguably a group can act in a decentralized fashion, without its action having to be intentional (i.e., goal oriented). My point is that individuals, taken as a group, do not necessarily have more intentionality than the environment they are part of, and so, under some descriptions, individuals can be viewed as acting on their environment in the same way that their environment is acting upon them. In chapter 2, I was interested in symmetrical relations between organisms and their environment. Symmetry can be broken again if the group of agents organizes – e.g., as a unified community (e.g., a urban public park city committee), or a social institution (e.g., the Legislator). The actions of that group can then be treated as that of a single individual. In that case, the asymmetry (i.e., the idea that one direction in the relation will be intentional, and the other not) between the group and its environment might be recovered.

Entities in a complex system will exist as individual entities, as groups of entities, and as organized groups of entities. Moving across the different scales of a complex system will involve treating entities differently; either as unified and equipped with the ability to take goal directed actions, or as disunified and lacking the ability to take goal directed actions. Across and within levels, there will be many relations possible (see fig. 2). In figure 2, the symmetrical relation discussed in chapter 2 is represented as the relation between "A" and "e" (A←→e). "A" represents an environment within a group of environments "E" made of different environments "A" and "B" (e.g., one of many ant nests). In turn, "e" represents an environment made of individuals 'a' and 'b' (e.g., individual ants in their nests). The relation between organisms and their environment discussed in chapter 2 is a relation "across scale" between "A" and "e" (A←→e, figure 2).

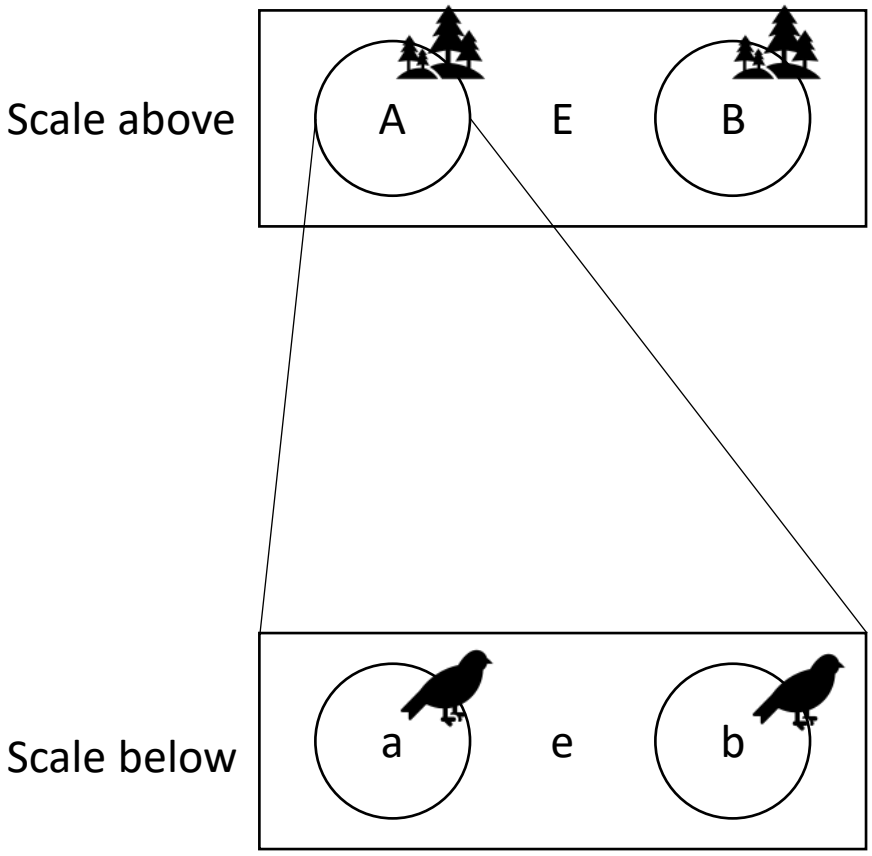


Figure 2. Relations within and across scales in a 2-level system. Circles represent units at the scale of interest, and rectangles represent environments including those units. Capital letters correspond to higher scale units and environments. The schematic shows the environment 'A' (e.g., a forest) within which one can find individuals 'a' and 'b' (e.g., birds), which grouped together form the population in their environment 'e'. The environment 'e' can itself be a unit 'A' at the higher scale. There exist relations within and across scales. Symmetrical relations can obtain across scales when considering the ensemble at the scale below (e.g., A←→e). Asymmetrical relations across scales are between an individual (or an ensemble treated as such) and the ensemble it is part of at the level above (or same level) (e.g., a→e and a→A). The key point is that an agent that we assume is endowed with intentionality will be able to enter in an asymmetrical relation with the environment (e.g., an individual bird "a" influencing its environment "e=A" a→e). Symmetrical relations are also possible between a group of organisms in their environment (e.g., "a" and "b" within "e") and that same environment when envisaged at the scale above (e.g., "A"). Symmetrical relations are characterised by an absence of intentional action on both sides, and allow us to account for self-organized phenomena like desire paths without postulating intentionality.

Between levels, other relations exist, such as that between a single individual in a group and her environment at the level above (e.g., a→A in figure 2). Such a relation does not need to be symmetrical, for in such a relation an individual can interact asymmetrically with her environment by imposing environmental transformations that suits the individual's needs and by refusing immediate feedback from the environment. Such an asymmetrical reading of the relation between an individual and her environment is useful when trying to account for person level phenomena (e.g., when asking questions about developmental processes of a single individual such as neuroplasticity in relation to the individual's cultural embedding). Then, by scaling down on that same individual, and by focussing on existing populations within that individual (e.g., bacteria

populations) and on the individual as being defined as a set of such relations, one can then return to a symmetrical perspective that allows one to inquire on subpersonal phenomena (e.g., the microbiome on sleep patterns of the individual).

Considering the individual as a set of symmetrical and asymmetrical relations is central to the selection of efficacious interventions in Psychiatry. If one assumes that the problem at hand involves an asymmetrical relation (e.g., the relations between neurodevelopmental trajectories and the person's social embedding), one might opt for an intervention centered on the intentionality of the actors involved in the problem (e.g., psychotherapy). In turn, if one assumes that the problem rests on a symmetrical relation, one might prefer an intervention that involves actors operating at the subpersonal level (e.g., pharmacotherapy). To get a full picture of the person's situation, one will have to scale up or down and across and within scales to integrate other factors into the diagnosis, so as to find the optimal management strategy. Within such a multiscale perspective, inquiring about any component of the entire system requires one to start by identifying the within and between scale relations that characterize the person. In adopting a person centric approach to psychiatry, for instance, a clinician might want to start by inquiring about the symmetry breaking relation between the person and her environment (e.g., the way individual life choice and experience contributes to marginalizing an individual in each cultural environment, and how the marginalization may contribute to sustaining a given affliction).

I recognize that all this is rather vague and general, but the point here is to stress the importance of moving across scales in our scientific or clinical inquiries to make sure that we are operating within the relations in the biosphere that correspond to the problem at hand. Symmetrical and asymmetrical thinking is at the core of the reasoning under a multiscale model (e.g., the biopsychosocial model), even when adopting a person-centric approach (e.g., the person centric approach in psychiatry). Knowing when one is dealing with a symmetrical or asymmetrical problem, that is, a problem for which the solution will be found by intervening on intentional or nonintentional relations is central to working within a multiscale perspective, such as that which we propose in the following chapter.

Chapter 3: Evolution, Culture and Computation in Psychiatry: An integrative perspective

Introduction to chapter 3

Chapter 3 is titled *Evolution, Culture and Computation in Psychiatry: An integrative perspective*. Explicitly, chapter 3 details the problem that motivates this dissertation, which could be summarized as the problem of a lack of a mechanistic understanding of the biopsychosocial model in psychiatry (e.g., in figure 1). To remedy that problem, chapter 3 proposes the integration of the evolutionary, cultural, and computational approaches to psychiatry. These are all multiscale approaches grounded in principled, functional, and mechanistic understandings of mental disorders. Evolutionary, cultural and computational approach are thus all interesting candidates for implementing the biopsychosocial model. However, each approach comes with its own disciplinary commitments, which appear to limit its ability to cover the full range of processes involved in the biopsychosocial model. We call this the problem of disciplinary boundaries.

The integrative model of chapter 3 is meant as a solution to the problem of disciplinary boundaries. As mentioned in the introduction of this dissertation, the model we propose in chapter 3 should be viewed (i) as the prototype of an integrative scientific ontology for psychiatry (i.e., a conceptual model that shapes education and practice based on principles of evolutionary, cultural, and computational psychiatry); (ii) as a framework to organize seemingly disparate knowledge that all matter for understanding mental disorders (e.g., by providing a lingua franca to talk about phenomena across scales, which is one of the roles of computational psychiatry); and (iii) as a practical guide to orient the way clinicians engage their client (e.g., by clarifying the different thinking patterns one can employ to make sense of mental disorders as entities).

Implicitly, Chapter 3 functions as the targeted literature review of this dissertation. This literature also fosters the simulation we present in chapter 4, which is an example of the sort of research output one can produce when operating under the model, or ontology, we turn to next. Chapter 3 also shows how active inference and the free-energy principle to model the relation between clinical interactions and institutions.

1 Introduction

1.1 The problem of disciplinary boundaries

Most of contemporary psychiatry assumes that gene–environment interactions over the course of developmental trajectories contribute to the aetiology of mental disorders (Adams et al., 2016). These trajectories depend on processes at multiple levels, including epigenetic, neurophysiological, behaviours and interpersonal interactions, which are embedded in larger systemic social contexts. Our currently limited knowledge about such interactions is a challenge for efforts to ground diagnostic nosology and clinical practice in a mechanistic understanding of the relations between multiple levels that constitute the complex pathways to mental disorders (Henriques, 2011). The aim of this chapter is to advance an integrative perspective that bridges three theoretical domains in psychiatry, which taken together, promise a mechanistic[10] understanding of the systemic processes and trajectories that underwrite psychopathology: evolutionary psychiatry, cultural psychiatry, and computational psychiatry.

Over the last 40 years, evolutionary, cultural, and computational psychiatry has each developed theoretical, empirical, and clinical approaches to psychopathology. Although representing different conceptual models and research methodologies, all three approaches aim to advance non-reductionist, mechanistic, and multilevel account of the pathways to mental disorders. As the name suggests, *evolutionary psychiatry* endeavours to explain mental disorders in terms of the evolutionary and genetic origins of the phenotypic traits (Stevens & Price, 2015). *Cultural psychiatry* emphasizes the role of culturally mediated social practices in development and the circular causality between illness behaviour and social context (Tseng, 2001). Finally, the emerging field of *computational psychiatry* studies failures in decision-making and dysfunctional behaviour using multi-level computational models (Friston et al., 2014).

Despite some recent exceptions (Badcock et al., 2017; Constant et al., 2021), each approach has remained largely siloed. This lack of dialogue results from institutional and conceptual difficulties in crossing disciplinary boundaries (Kirmayer, Worthman, & Kitayama, 2020b). Disciplinary boundaries are the consequence of particular research histories and traditions but also reflect specific scientific ontologies

[10] In this paper, we use a folk concept of mechanism — of the sort that any typically trained psychiatrist would have in mind in clinical case formulation. In this context, a mechanistic approach may be loosely defined as one that analyses the causal processes that produce a given (psychiatric) outcome, through reference to constituent components and their interactions. There is an important debate in the philosophical literature about the precise nature of mechanism (Machamer et al., 2000; Nicholson, 2012; Rosenberg, 2020). Our previous work on the embodied and situated human brain is aligned with a neo-mechanistic perspective in the philosophy of science (e.g., Bechtel, 2009; Craver, 2006; Glennan et al., 2021; Rosenberg, 2018), which explains the properties, functions, and behavior of a system by examining the properties and activities of its various subsystems and their interactions. Here, a mechanism can be described as a structure (or a stabilized process) within a system that performs a function via its component parts, their various operations, and their organisation, thereby contributing to global function in one or more ways. According to previous work on the hierarchically mechanistic mind (Badcock, Friston, Ramstead, et al., 2019) consistent with the present approach, the human phenotype is produced by causal mechanisms that span both spatial scales (e.g., genes, cells, tissues, organs, the body, and the broader social and physical environment), as well as temporal scales (ranging from evolutionary/intergenerational processes, through to developmental influences, and mechanisms that operate in real-time biopsychosocial contexts) (Kirmayer, Worthman, & Kitayama, 2020). In short, this multilevel theory describes human phenotypes in terms of the biopsychosocial processes that operate within and across different spatiotemporal scales, and in this sense, it is both mechanistic and hierarchical.

(Latour, 2000). Ontologies underwrite research agendas(Hacking, 1995), which reflect researchers' beliefs about what questions science should address and what kinds of answers are satisfying (Kuhn, 1962), and that lead researchers to operate under different 'thought styles' (Fleck, 1979). Disciplinary ontologies require that researchers become skilled at using specific methods, which render measurable and ontologically "real" or conceivable certain dimensions of the object of inquiry (Pickering, 1995). By the same token, due to constraints of time and resources, commitments to disciplinary ontologies also limit researchers' skills and impede the study of certain dimensions of phenomena and may make them invisible or even inconceivable. The result then is progress on some fronts but lack of attention to other, possibly crucial, facets or dimensions. This effect of disciplinary ontologies is especially concerning in the context of psychiatry, which is concerned with human problems that clearly involve multiple processes that affect physiology, behaviour, and experience (Kendler et al., 2020). Advancing an integrative perspective, requires some way to move beyond these disciplinary blinders. We propose that unifying cultural, evolutionary and computational psychiatry can enable significant strides towards an integrative view.

1.2 The scope of the integrative perspective

This chapter starts with an overview that lays out some assumptions and methodological strategies employed in evolutionary, cultural, and computational psychiatry (§2). We will not discuss evolutionary, cultural or computational psychiatry in their entirety. Rather, we focus on key aspects of these approaches—mainly modes of reasoning about mental disorders—that could be merged through an interdisciplinary way of thinking about mental disorders.

Key aspects of evolutionary psychiatry

With respect to evolutionary psychiatry, we will focus on adaptationist reasoning about pathological mental traits, which can be distinguished from population genetics thinking (Keller & Miller, 2006). Adaptationist reasoning in evolutionary psychiatry emphasizes the role of natural selection when making sense of mental traits observed in clinical settings (Grunspan et al., 2018); population genetics thinking may be viewed as a research-driven attempt at explaining changes in the genetic makeup of a population and the preservation of alleles that contribute to certain mental disorders (Keller & Miller, 2006). The distinction between population genetics thinking and what we call adaptationist thinking can be framed more generally in terms of what some historians have identified as the distinction between the modern synthesis and the ethological perspective (Adriaens & De Block, 2010). This division is interesting in that it carves out two interconnected questions about mental disorders that are approached with distinct reasoning patterns. The ethological perspective asks, "How can we understand mental disorders as traits that have evolved in humans and other animal species to serve certain functions?" and seeks answers based on the relation between phenotype (e.g., behaviour) survival value and fitness. In turn, the modern synthesis perspective asks, "How can we explain the preservation of alleles underlying mental disorders in a population?" and seeks answers based on a wide array of evolutionary mechanisms, including, though not limited to, the logic of survival and fitness under natural selection (e.g., drift, mutations, gene flow). The extension of the modern synthesis—the extended evolutionary synthesis (Laland et al., 2015) —suggests supplementing the mechanisms of evolution with channels of

inheritance and processes that are external to the organism (e.g., cultural inheritance, niche construction and development). The adaptationist rationale—on which we focus—can be assimilated to the ethological perspective.

Key aspects of cultural psychiatry

With respect to cultural psychiatry, we will focus on how cultural context may shape mental disorders through a variety of intra- and interpersonal feedback loops, including what Hacking has termed "the looping effect of human kinds" (Hacking, 1995). Cultural psychiatry studies the ways in which culture and social context shape the aetiology (causes), phenomenology (experience), clinical presentation (expression), and trajectory of mental disorders (Kirmayer & Ryder, 2016). This includes the person's own modes of self-construal and the responses of others, which draw from cultural narratives, models and metaphors. Taken together these constitute the ontology of a mental disorder. Although this will not be our focus here, it is important to note that cultural psychiatry also leverages the notion of culture to orient clinical assessment, treatment, and prevention (e.g., situating illness experience in its social and cultural context to identify the significance of cultural expressions of distress and their impact on the course and outcome of mental health problems)(Kirmayer, 2005). Cultural psychiatry also emphasizes self-reflexive practice, through studies that reveal the cultural assumptions the institutions of psychiatry itself (e.g., ethnocentric biases) that may affect mental health research and clinical practice as well as illness experience (Kirmayer, 2018; Young & Breslau, 2016).

Key aspects of computational psychiatry

With respect to computational psychiatry, we will focus on the rationale of modelling psychiatry. Computational psychiatry involves the use of algorithmic methods to model and analyse clinical and behavioural data (Gauld et al., 2021). This includes two broad, though interrelated lines of work in computational psychiatry: (i) *data-driven* computational psychiatry, involving the use of artificial intelligence and machine learning with large datasets ("big data") to develop more precise characterizations of patients that have some predictive validity in relation to treatment response and course of illness; and (ii) *theory-driven* computational modeling, which develops biologically plausible accounts of neural processing that can explain particular forms of psychopathology (Huys et al., 2016). The focus here will be on the latter approach, which aims to understand the mechanisms of psychiatric disorders by constructing computational models.

Our proposed integration of cultural, evolutionary and computational psychiatry aims to show how adaptationist thinking and the social-cultural notion of looping effects can be integrated using the methods of modelling psychiatry. To illustrate the potential of this integration, we describe a generic model for the study of mental disorders that inherits principles of evolutionary and cultural psychiatry (§3). The hope is that the resulting Evolutionary Computational Cultural (ECC) model will exemplify the interdisciplinary approach we advocate. The end of part three illustrates an application of this model using the clinical example of Major Depressive Disorder (MDD).

2 Evolution, Culture and Computation in Psychiatry

The difference in disciplinary ontologies poses a central theoretical challenge for collaboration among evolutionary, cultural and computational psychiatry. How can we think through the ideas of the evolutionary approach in computational terms; of cultural ideas in evolutionary terms; or computational ideas in cultural terms? What do we need to know to map one theoretical construct onto the other and what concepts and relations require special attention? This process of inter-theoretic mapping needs to start with a general understanding of principles employed in evolutionary, cultural and computational psychiatry. We will consider adaptationist thinking, the ontology of mental disorders, and modelling psychiatry.

2.1 Evolutionary psychiatry

Medicine often employs functional models of health and diseases based on principles of human physiology. These models indicate how the body is supposed to function. Pathology can then be identified as a disruption or impairment of this function (Boorse, 1982). For instance, we assume that the heart is designed to pump blood; and this is why, no matter the cause, congestive heart failure may be confidently described as a malfunction (Nesse, 2007; Schwartz, 2007). Although efforts have been made to define mental dysfunction in a similar way (Wakefield, 1992b, 2005), this effort has been impeded by the fact that the human mind has multiple functions that depend on adaptive context. Attempts to characterize brain function are intensively debated (Montague et al., 2012). One consequence of this lack of clarity about the functions of mind and brain is difficulty in distinguishing between disorders and protective responses (Nesse, 2007). For instance, we know that congestive heart disease is a disorder and that fever is a protective response, because the former can be said to result from a failure of a function of the heart (e.g., pumping blood), whereas the latter reflects a functional biological response to infection (Nesse, 2007).

Evolutionary psychiatry has sought to address this limitation by exploring plausible functions of mind and brain against the backdrop of human evolution. By applying the principles of evolutionary biology and psychology, evolutionary psychiatry aims to provide a basis to distinguish normal and pathological mental functioning, based on the notion of adaptive fitness (Nesse, 1999). This leads to a view of mental disorders as 'harmful dysfunctions' (Wakefield, 1992a). As will be detailed below, in the account of mental disorders as ‘harmful dysfunctions’, the dysfunction refers to the functional aspect of the proximal mechanism (e.g., regulation of dopamine signalling), whereas the failure is defined in terms of discrepancies with respect to the way that mechanism ought to function from an evolutionary point of view (e.g., regulation sufficient to enable an adaptive response to the environment). In turn, the ‘harmful’ component refers to value-laden terms that are often qualifiers of the disorder (e.g., autistic individuals’ “lack of motivation”). Note that for Wakefield, the distinction between the dysfunctional and the harmful brought with it putative problems that would need to be addressed in a more fully developed version of his approach.

The problems that surround Wakefield’s concept of mental disorder are at least twofold (Faucher & Forest, 2021). First, there is the problem of identifying the evolutionary adaptive process against which the

dysfunctional mechanism can be evaluated: this has been termed “the problem of evolutionary function.” Second, there is the problem of the scientific validity of the notion of ‘harmful’, which is generally recognized to be, at least partially, socially and historically contingent. Indeed, according to the view of harmful dysfunction theory, although value-laden qualifiers are an essential part of the definition of mental disorders, the study of their functional role is difficult to assimilate to a purely evolutionary view. Yet, as argued by cultural psychiatry, unpacking the meaning of “harm” and other evaluative qualifiers is essential since psychiatric disorders are both biological and social constructs always occur in particular cultural contexts. This article will focus on the latter problem. In section 2.2, integrating a cultural approach will allow us to address this problem by providing a more complete view of the mechanisms of mental disorders that explicitly incorporates humanly constructed contexts and corresponding social interactions.

2.1.1 Defining mental disorders with proximate and ultimate thinking

Evolutionary psychiatry proposes a research heuristic for the study of mental ill-health, organized around the question of “why did evolution leave us with traits that make us vulnerable to mental disorders?” (Graves et al., 2016). This framework integrates proximate (e.g., developmental) and ultimate (i.e., evolutionary) levels of causation when defining mental disorders (for a summary see: Bateson & Laland, 2013). Sciences that study proximal mechanisms typically answer questions of the form ‘how does it work?’, (e.g., ‘how does experience-dependent neuroplasticity operate?’), whereas sciences that study ultimate causes answer evolutionary questions of the form ‘why does it work?’; (e.g., ‘why has experience-dependent plasticity been preserved throughout human evolutionary history?’) (Kenrick, 2001).

Evolutionary psychiatry defines mental disorders as dysfunctions of adaptive systems (or consequences of adaptive systems that are maladaptive in a new niche or context), and explains disorders in terms of vulnerabilities aggravated by developmental demands. Note, however, that this type of explanation remains controversial (Varga, 2012). Some mental disorders have been viewed as adaptive dysfunctions, that is, as adaptations *per se* (e.g., psychopathy as an adaptive strategy from a game theoretic point of view (Murphy, 2005)). In this review, we will not pursue the view of mental disorders as adaptive dysfunctions. Rather, we will focus on explanations in terms of aggravated vulnerabilities. The integration of proximate and ultimate causes allows evolutionary psychiatry to study the impact of evolutionary pathways on the nature of mental disorders and their expression over the lifespan. The proximate part of this view describes the workings of the specific mechanisms underlying the development of pathology and their expression in symptomatology, suffering or functional impairment. Conversely, the explanation in terms of ultimate causes involves relationships between mechanisms and traits (and their associated vulnerabilities) that are conserved over evolutionary history (Nesse, 2017). In short, integrating proximate and ultimate causes allows evolutionary psychiatry to explain psychiatric conditions from the point of view of vulnerabilities stemming from phylogenetically old traits (Del Giudice, 2014).

Proximate and ultimate thinking in psychiatry tends to operate under two interrelated modes of evolutionary thinking: adaptationist and population genetics. Of course, the distinction between evolutionary influences that constitute proximate and ultimate causes is made for epistemological reasons. A more fine-grained assessment of causality would consider phenomena across multiple spatiotemporal scales, ranging from biochemical to evolutionary, including the scales of individual developmental trajectories and of the coevolution of the human brain and our cultural niches (Kirmayer et al., 2020a). The strategy of dividing causality into proximate and ultimate causes allows us to distinguish phenomena about which we can meaningfully ask questions like "Why has it evolved to work that way?" from phenomena about which we would better ask "How does it work?". For instance, ultimate causes may capture phenomena that unfold on a historical timescale or longer for which answers to "how" questions will likely remain uncertain (e.g., "What were the exact mechanisms at play in the evolution of this population?"), and for which the response to a "Why" question may be preferred (e.g., "What principles of evolution can explain why this feature might emerge?").

2.1.2 Adaptationist thinking

One popular strategy for the study of evolutionary pathways to mental disorders is the adaptationist approach (Troisi, 2006), which relies on the notion that evolution selects for functions that improve reproductive success. Evolution favours the replication of variations that lead to reproductive success (fitness). Since differential reproductive success is correlated with being adapted to environmental stressors, the genetic material passed onto offspring should lead to phenotypic traits that will be adapted, or well 'designed', to respond to these stressors in offspring (Houston et al., 1999). As applied in evolutionary psychiatry, adaptationism relies on the idea that vulnerabilities are shaped by Darwinian selection. Typically, it is not that natural selection selects 'for' disorders (e.g., viewing disorders as affording some fitness advantages) (Nettle, 2004). Rather, ultimate causes must be viewed as shaping genetic traits that may be expressed as suboptimal traits or vulnerabilities under certain proximate, developmental conditions (Nesse, 2017). Put another way, the maintenance of "any suboptimality [or vulnerability] of a part is explained as its contribution to the best possible design for the whole" (Gould & Lewontin, 1979, p. 586). Again, the question is not "how do genes that predispose to a mental disorder provide a selective advantage?" (Nesse, 2011), the answer to which would explain why mental disorders exist; nor is the question directly, "how do genes that predispose to a mental disorder persist?", the answer to which would explain why some mental disorders continue to exist. Rather, the question is "why are we vulnerable to some mental disorders?", the answer to which explains the clinical presentation of the mental disorder in the current context. This is important because explanations in psychiatry should be explanations of mental disorders, not only explanations of their underlying biology. As we will see with cultural psychiatry, mental disorders are entities configured at the level of human agency and subjectivity. Inquiring how aspects of a person's biology make that person vulnerable to a mental disorder is usually more immediately relevant to clinical practice than exploring the evolutionary origin of that biology.

Darwinian rationales have been used to explain different pathways to mental disorders in terms of the maintenance of vulnerabilities in human evolutionary history (ultimate cause) enabled by developmental context (proximate cause). Box 1 summarizes some of the popular rationales in adaptationist accounts of

medicine in general. Darwinian rationales have been applied to explaining mental disorders such as anxiety, phobic, delusional, stress-related and depressive disorders among other mental health problems (Durisko et al., 2015; Karasewich & Kuhlmeier, 2020; Troisi, 2020; Tsou, 2021). Importantly, all of these approaches assume the embeddedness of the individual in a larger systemic context. For instance, following a Darwinian rationale, the social risk hypothesis of depression (Allen & Badcock, 2003, 2006; Badcock et al., 2017) argues that normative symptoms of depression—triggered by social uncertainty—form an adaptive biobehavioral strategy that might have been selected to ensure the re-stabilization of individuals' social networks. Here, depression is thought to reduce socio-environmental volatility via three broad classes of action: it increases an individual's cognitive sensitivity to social risks; it reduces her propensity to engage in social behaviours with uncertain outcomes; and it promotes social signalling behaviours to elicit interpersonal support and defuse competitive encounters (e.g., reassurance seeking). When these responses fail to alleviate social stress (e.g., signalling fails to increase interpersonal support), depressive symptoms endure, and the individual can spiral into more severe and persistent distress that is recognized as clinical depression. To account for the prevalence of depression in a given population, from an epidemiological perspective, one could couple the social risk hypothesis with an evolutionary mismatch rationale (see Box 1) to explain why depression may increase in a society in which people tend to have sparse human social networks.

Box 1. Adaptationist Explanations for Psychopathology

Mismatch: Vulnerabilities may emerge from differential rates in evolution that generate mismatches between the cultural developmental environment and evolutionarily old dispositions (e.g., disordered eating patterns leading to obesity, because of humans' tendency to seek energy-rich, sugary and fatty foods that were scarce in our ancestors' environment but that are now abundant) (Raubenheimer et al., 2015). A mismatch happens when the rate of change of environmental stressors exceeds the rate of change of individuals' adaptation. Depending on the scale at which the mechanism of adaptation lags behind, a mismatch will either be defined as *developmental* – i.e., a body-environment mismatch (Bateson, 2001); or *evolutionary* – i.e., a genotype-environment mismatch (Bourrat & Griffiths, 2021; Riggs, 1993). Developmental mismatches are assumed to impair realized fitness (i.e., individuals' reproductive success), whereas evolutionary mismatches are assumed to impair the ability to achieve expected fitness (i.e., the sum of reproductive success weighted by fitness across all possible environments).

Constraints: Constraints on selection arise when the cost of adapting a vulnerability through natural selection is higher than the cost of preserving that vulnerability in the population. For instance, the cost of delivering human infants through the pelvis, although painful and often dangerous, does not outweigh the cost of reengineering the birth canal (Nesse, 2017). 'Rule of thumb' logical reasoning outweighs its cost in terms of logical errors (Mercier & Sperber, 2017); and Huntington's disease has a limited cost since its symptoms do not appear before the age of child-bearing (Nesse, 2002).

Trade-offs: Trade-offs also favour the selection of vulnerabilities understood as defenses, according to 'smoke detector' explanations (Nesse, 2001). Smoke detector explanations apply in cases where it is more cost efficient to select for genes that result in traits likely to trigger false alarms than to fail to detect threat (e.g., predator, or fatal pathogen). For instance, acute sensitivity to anxiety provoking situations increases

the success of fight or flight responses (and thereby contributes to reproductive success), but it increases vulnerability to social anxiety disorders. Similarly, fever is a defense against infection (Kluger, 1979) but it may increase to the point of causing seizures.

2.1.3 Limits and prospects of adaptationist rationales

Adaptationist accounts explain mental disorders in terms of the vulnerabilities of systems that evolved to serve an adaptive function (e.g., depressive symptoms are an adaptive vulnerability whose function is to reconsolidate social networks but that can spiral into maladaptive responses). This makes an explicit link between normal functioning and pathology and provides a rationale for research with animal models that involve similar biobehavioral systems (Kirmayer & Crafa, 2014); hence the ties of adaptationist thinking with the ethological perspective. Adaptationism has been critiqued, however, on methodological and conceptual grounds (Tsou, 2021), among others, on the fact that traits may persist and lead to vulnerabilities through processes other than selection (Gould, 1991). Indeed, there are many cases that cannot be explained solely based on Darwinian thinking. For instance, disorders such as schizophrenia, bipolar disorder, eating disorders, and obsessive-compulsive disorder are known to impair reproductive success (Keller & Miller, 2006). All things being equal in the world of natural selection, genetic variants predisposing individuals to such disorders (e.g., genetic vulnerabilities) should have been eliminated from the gene pool long ago. To explain pathways to mental disorders based on traits that have no obvious adaptive value, evolutionary accounts of mental disorders can go beyond the adaptationist narrative by appealing to other population-level phenomena.

Explanations based on population genetic thinking provide a complement to Darwinian explanations (for a review see: Keller & Miller, 2006). For instance, processes of *balancing selection* can maintain multiple variations of alleles in the same gene (i.e., polymorphism) whose net fitness effects balance each other out, depending on the genetic or environmental context (Zhang & Hill, 2005). Balancing selection requires that all the alleles involved have roughly equivalent fitness, and that some mechanisms countered the normal loss of these alleles due to drift. A good example of a balancing selection process is frequency-dependent selection, where the fitness of some unit (e.g., allele AA) or trait depends on its frequency in a population (e.g., the hawk-dove situation (Sigmund & Nowak, 1999)). Frequency dependent selection might explain the maintenance of allelic susceptibility to psychopathy, as people with psychopathy would gain a fitness advantage in a population where the allele is rare and becomes disadvantageous when frequent because of anti-cheater vigilance (Mealey, 1995; Nettle, 2004, 2006). Like adaptationist rationales, rationales from population genetics explain the persistence of dysfunctional genetic variations (e.g., vulnerabilities to illness) that would normally impair evolutionary success. This provides evolutionary psychiatry with a functional model of mental health and disease based on biological principles. It is important to note that there are many other population genetics models that can explain the persistence of harmful variations (Keller & Miller, 2006). The example of balancing selection is introduced here to warn against overly simplistic adaptationist stories, which are often difficult to test. That said, adaptationist accounts can provide satisfying explanations for some mental disorders. Crucially, adaptationist rationales point to the likelihood that many mental disorders are based on otherwise adaptive functions (Grunspan et al., 2018). These rationales can lead to rethinking medicalization

or conventional psychiatric nosology by acknowledging the close links between adaptive strategies and pathology (Troisi, 2005).

There are also limitations to the adaptationist approach that are external to it. As the logic of evolutionary biology goes, proximate causes acquire explanatory value in so far as they relate to ultimate causes, which are located in evolutionary history. However, in many instances, this history refers to the emergence of human beings in an evolutionary environment of adaptation quite different from our current environments. Evolutionary explanations either appeal to vulnerabilities that arose because of this evolutionary history or focus on discrepancies between past environments, to which we were well adapted, and current contexts, which pose new challenges (cf. Box 1). New challenges in current contexts, however, are dependent upon sociocultural features like cultural practices, values and social institutions, whose causal contribution to mental health should be considered (Kirmayer & Young, 1999). Moreover, humans have been co-evolving with our socially constructed environments for millennia (Kirmayer et al., 2020a). Thus, half of the story is missing here. As we will see next, cultural psychiatry provides a concept of mental health consistent with evolutionary thinking, which can provide a mechanistic account of the social systemic embedding of mental health and illness.

2.2 Cultural psychiatry

Cultural psychiatry acknowledges the influence of multiple processes in establishing the boundaries between the normal and the pathological in biomedical science and clinical practice (Kirmayer & Crafa, 2014). However, it insists that any perspective must acknowledge context dependence; that is, the influence of socio-normativity of the local cultural contexts. This is crucial to produce definitions of mental disorders that have a grip on clinical practice. Moreover, cultural psychiatry argues that evolutionary history itself is shaped by current cultural concerns and dominant ideologies that may obscure the nature and range of human functioning in health and illness (Canter, 2012; Rose & Rose, 2010). Accordingly, for cultural psychiatry, an evolutionary perspective must consider the social normativity that underlies the use of evolutionary principles to define the normal (functional) and the pathological (dysfunctional) (e.g., the manner in which values of a local ethnomedical practice shapes illness experience and thereby themselves move the boundaries of the normal and the pathological (Kirmayer & Young, 1999; Murphy & Woolfolk, 2000).

Cultural psychiatry does not endorse a radical social relativism, which would discount any effort to recognize mental disorders across cultures. Mental disorders are not simply social constructions; they are fundamentally biological. But cultural psychiatry insists that human (neuro)biology is itself fundamentally social — neurodevelopment and adult functioning involves the embedding of the individual in a socially constructed niche and larger interactional systems that are configured by cultural knowledge and practices (Kirmayer, 2006). Recent human evolution has involved cultural-biological coevolution, so that even our thinking about mental disorders in evolutionary terms must engage with the impact of humanly constructed worlds on the structure and function of our brains. Moreover, changes in these social and cultural systems happen faster than evolutionary changes creating potential discrepancies between functional systems and current adaptive demands. The key questions for cultural psychiatry then are not only those that relate to the way in which the

social world shapes the experience, definition of, and response to mental disorders, but equally how social contexts and interactions contribute to the underlying mechanisms and developmental trajectories of disorders: that is, how and when mental disorders are constituted by processes that reflect their social systemic embedding.

It is hard to see how one could disagree with the holistic view of mental health proposed by cultural psychiatry. Yet, historically, these claims have been given mostly lip service, as bio-reductionism still appears to run deep in psychiatry. To understand the project of cultural psychiatry, we must take a short glance at the recent history of psychiatry and the concept of mental disorder it has employed.

2.2.1 Historical overview of bio-reductionism

The operationalization of diagnostic categories ushered in by DSM-III in 1980 aimed to provide a taxonomy useful for clinical assessment that could also guide research aimed at identifying discrete disorders, each with its own aetiology, mechanisms and symptoms (First, 2012). Categorical approaches were born from a "biomedical" approach to research and practice that focused on the proximal, biological factors at play and their associated phenotypes (Compton & Guze, 1995; Mayes & Horwitz, 2005). The categorical approach of the DSM-III and its successors emerged against the background of already ongoing arguments for a broader *biopsychosocial approach* to assessment (Engel, 1981; Guillemin & Barnard, 2015). On the biopsychosocial view, the illness must be understood in terms of a multilevel hierarchy from molecules to behaviour. This affords a conceptual space that accommodates clinical observations in the real-world contexts of disorders (Bolton & Gillett, 2019). However, the hope of characterizing disorders in terms of underlying (biological) mechanisms and the lack of appreciation of the causal effects of systemic social processes has undercut integrative approaches(e.g., Ghaemi, 2009, 2010).

The current Research Domain Criteria (RDoC) developed by the United States National Institute of Mental Health reflects the emphasis on biological correlates, as it doubles down on neuroscientific research, with the hope of formulating disorders in terms of their (mostly neural) phenotypes and/or measurable (neuro)biological traits (Insel, 2014). Despite the integration of behavioural and phenomenological (e.g., through self-reports) units of analysis, the RDoC framework remains largely bioreductionist (Kirmayer & Crafa, 2014). In emphasizing biological research, the RDoC relies heavily on evidence derived from animal models. Unfortunately, we have no animal models of many distinctive components of human experiences relevant to mental health and illness, such as narrativity, morality, racism, political violence (Paris & Kirmayer, 2016). Reductionism thus is bound to operate with a stripped-down biology that emphasizes brain circuitry over psychological functions and systemic social processes. This makes it difficult for psychiatry to advance its goal of a mechanistic understanding of all the components that make up the gene-brain-person-environment pathway to explain mental disorders. Importantly, RDoC criteria do not have immediate clinical utility, as their role has been mostly to motivate "omics" research. Omics can lead to novel treatment that target the biological components of mental disorders (Morris et al., 2022), but, as the currently struggles of the psychiatric institution with the "translational gap" suggests – i.e., the difficulty in translating scientific models into efficacious clinical models (Seigle et al., 2019) – such research has been largely unsuccessful. Cultural

psychiatry seeks to move towards a concept of mental disorder that remains mechanistic and functional while accommodating culture and context.

2.2.2 Towards a non-reductionist concept of mental disorder

The concept of disorder in psychiatry refers to behavioural patterns that cause psychological distress and functional impairment, and only indirectly to the failure of biological mechanisms. It describes a situation configured at subjective, phenomenological, psychological and social systemic levels (Kirmayer & Young, 1999). Mental disorders are inherently value-laden and shaped by socio-normative causes — e.g., the way we identify the harm resulting from mental ill-health — as much as they are caused by biological causes.

In considering distinctions between health and pathology, cultural psychiatry raises an additional difficulty: namely, giving a scientific account of 'harmful'. We need to identify and test the mechanisms by which judgments themselves, understood as objects of language, become consequential for individuals' functioning, well-being, social status, etc. Institutional discourse shapes illness experience, which means that we need a functional account of how individual and institutional discourse influence the mind, and how the mind comes to affect institutions. In the notion of 'Harmful dysfunction', the harmful and the dysfunctional must be given equal scientific consideration.

Accordingly, cultural psychiatry defines mental disorders: (i) *pragmatically*, as conditions treated by the discipline of psychiatry, or corresponding local healing practices; (ii) *normatively*, relative to the conceptions of the normal and the pathological given by local medical traditions and practices; and (iii) *ontologically*, as having bodily, psychological, or social systemic causes (Kirmayer, 2018). In employing cross-cultural and ethnographic methods, cultural psychiatry can work out the pragmatic and normative aspects of mental illnesses (e.g., assessing the manner in which individualism in Western culture impacts health and well-being (Eckersley, 2006, 2011; Kirmayer & Ban, 2013; Kirmayer & Bhugra, 2009). Here we focus on the ontology of mental disorders but recognize the fact that the category of pathology is a moving target influenced by language and culture. Although this remains a challenge, cultural psychiatry captures the moving aspect of ontology using the theory of the looping effects of human kinds, developed by Hacking (1995, 2000). We believe that one can leverage the mechanics of looping effects of human kinds to think about a scientific study of the 'harmful' in Wakefield's concept of mental disorder.

Kinds are epistemological notions that refer to conceptual classes used to classify, sort or discriminate different objects (Hacking, 1995). Natural kinds, for instance, classify objects that undergo *efficient causality*, in the sense that when they are acted upon, those objects conserve the same set of properties. However, objects classified as humankinds, such as mental disorders, do not only undergo efficient causality; they undergo *practical causality*—that is, they change their behaviour by virtue of the act of being classified or labelled. This means that kinds are desirable, or undesirable to the people whose behaviour fall under their classification (Hacking, 1995). It is because they are value laden (that is, they depend on the values assigned to them through social practices) that humankinds are endowed with a causal power different from that of

natural kinds. For instance, if N is a natural kind, and Z is an object of the natural kind N, classifying Z as an element of N has no causal effect on Z (Hacking, 1995). For instance, if 'atom' is a natural kind, calling an 'atom' 'hydrogen' has no causal effect on hydrogen as an atom. What might change is the way the classifier would engage with hydrogen. The same applies to human kinds (e.g., if I call Denis 'autistic', it will change the way I engage with him). However, while classifying 'atom' as 'hydrogen' changes only the behaviour of the classifier, classifying Denis as autistic also changes Denis' behaviour. In contrast to the atom, Denis can become aware of his classification and may change his behaviour accordingly. Denis might make less effort, or lose motivation to engage socially because of self-perception and self-evaluation based on his understanding of the classificatory label, or because of his internalization of the stereotypes and social stigma applied by his social partners (Jaswal & Akhtar, 2018). These proximal interactional effects are, of course, embedded in larger systemic social processes and structures that are major determinants of health and illness (Alegría et al., 2018; McAllister et al., 2018).

Categories of mental disorders are about people and the criteria they are based on often reference behaviours that are value laden. In turn, our categories of people, their character and values are all culturally shaped (Hacking, 1985; Kirmayer, 2007). This leaves the ontology of any given mental disorder open to change as a function of local cultural changes in norms, conceptual categories, and epistemic practices. For instance, as diagnostic activity and treatments may recognize certain configurations of experience and affliction, clients may access new ways to interpret their experience, thereby yielding corresponding clinical presentations that reinforce the clinician's impression of the validity of the category (Kirmayer, 2018).

Looping effects may entail a shift from one locus to another, such as in cases of somatization, where the affliction may start as a social experience, and then become psychological, bodily, and then social again. Somatization is found across cultures (Kirmayer & Young, 1998) and appears to reflect basic psychophysiological processes that are shaped by culturally specific ways of life and modes of illness experience. These modes of illness experience are culturally patterned ways of expressing bodily and psychological afflictions that reflect cultural models (Kirmayer & Sartorius, 2007). Cultural models are stable discursive and expressive styles of illness experience encoded in individuals' cognitive schema, embodied practices, interpersonal interactions, discourses, and social institutions. It is these cultural models that lie at the interface of individuals and larger social world to mediate the looping interaction between somatic and emotional/psychological distress (Kirmayer & Ramstead, 2017). Looping effects in cultural models are promising candidates for a mechanistic account of the harmful, in Wakefield's definition of mental disorder as harmful dysfunction.

2.2.5 Prospects for an ecosocial model of mental health

Cultural models point to a concept of mental disorder that recognizes the causal power of social labelling of behaviour and experience as *harmful* and aligns more generally with the biopsychosocial approach that recognizes individual cognitive and adaptive processes are embedded in larger systemic social contexts (Bolton & Gillett, 2019). Cultural psychiatry situates the open-ended looping ontology of mental disorders in an ecosocial model of mental health (Kirmayer, 2015), which—much like recent multilevel approaches in

psychology (Badcock, 2012; Badcock, Friston, & Ramstead, 2019; Badcock, Friston, Ramstead, et al., 2019) and cognitive anthropology(Veissière et al., 2020)—assumes that humans are part of a hierarchically organized, dynamic social ecosystem that includes the brain, the body, and the social and physical environment (Hutchins, 2010). This means that psychopathological entities may involve dysfunctions not only in their subcomponents (e.g., neuroatypicalities; bodily impairment; dysfunctional social milieu), but in the system dynamics that bind these components together (Borsboom et al., 2019). These dynamics include feedback regulatory processes and mutually causal looping effects that can amplify or self-sustain a psychopathological state (Kirmayer, 2015).

The ecosocial model of mental health gives explicit attention to the systemic embedding of human biology and psychology by drawing links or loops between our self-descriptions (as ill or well) and interactions with the brain, body, and society. It encourages us to consider the multiple forms of systemic social process that give rise to human experience in sickness and in health. In so doing, cultural psychiatry aims to lay bare not only the constructs, norms and constraints that constitute mental disorder as a social reality, but also the cognitive and social interactional processes that may be aetiological factors, part of basic mechanisms of psychopathology, and determinants of illness course and outcome. The resultant models of pathology trace the circuits of the mind, which reside not only in the brain but in the social world. However, although well framed to advance an integrative approach, the ecosocial model of cultural psychiatry, in its current form, remains mainly a narrative description of the mechanisms at the interface between external levels of causation (e.g., socio-material systemic processes) and internal (e.g., brain-based) levels of causation, making it difficult to operationalize in empirically testable models (Smoller & Stein, 2018). To remedy this, we next consider ways to implement looping dynamics—that undergird the ecosocial model—within the formalism of computational psychiatry.

2.3 Computational psychiatry

As a domain of clinically applied research in psychiatry, computational psychiatry is primarily motivated by recognition of the shortcomings of current psychiatric nosology in providing diagnostic categories that predict treatment response and outcomes and that are linked to mechanistic explanations of disorder (Corlett & Fletcher, 2014). However, computational methods also allow us to build models of biological processes that are systemic—that is, they can model networks of many interconnected components and reveal the resulting dynamics. This has proved a powerful approach in systems biology at many levels and, in particular, in efforts to understand how embodied and embedded and extended neural networks can give rise to cognition, behaviour, and experience in health and illness.

In this section, we focus on an approach to theory driven modelling in psychiatry known as active inference (Da Costa et al., 2020; Friston, FitzGerald, et al., 2017; Parr et al., 2022). Active inference has been proposed as a general framework for understanding the computational processes that underlie cognition and adaptation, which essentially involve prediction of sensory inputs and the effects of actions. This approach understands mental disorders in terms of failure to infer or represent causes of sensations in the world based on Bayesian beliefs, and to act accordingly (Corlett & Fletcher, 2014).

Under active inference, mental disorders are defined and modelled in terms of a failure of cognitive functions such as (i) perceptual inference and (ii) adaptive behaviour as action planning. Within the terms of our current discussion, active inference can be viewed as seeking an explanation of proximate causes of dysfunctions, where dysfunctions should be understood as suboptimality of perception and action (Parr, Rees, et al., 2018). The question active inference asks is: Assuming that the brain operates optimally from a Bayesian point of view (i.e., it always performs Bayesian inference), how is it that the brain can generate suboptimal behaviour? This question is close in nature to that of evolutionary psychiatry: if natural selection optimizes organisms' adaptation, how is it that natural selection can generate suboptimal phenotypic traits (e.g., vulnerabilities)? In both cases, the answer is that suboptimality is the outcome of an optimization process that has 'gone wrong' given the developmental, environmental, or social-contextual conditions under which the maladaptation emerged. For evolutionary psychiatry, things can go wrong, for instance, because of a mismatch between the environment of evolutionary adaptation and contemporary social contexts. For computational psychiatry, the optimization process goes wrong when something happens to the priors involved in the cognitive machinery, because of lesions, autoimmune, neoplastic, infectious, or neurodevelopmental anomalies, alterations in neurochemical or neuromodulatory processes, or changes in brain circuitry that may be a result of environmental interactions and learning histories. However, drawing from the arguments of cultural psychiatry, this circuitry may involve systemic processes that extend beyond the brain. We will explore those processes in section 2.3.3 below.

2.3.1 Computational phenotypes

The theory of active inference allows one to produce computer models of pathological and healthy brain functions to study the effects of various kinds of interventions (mostly psychopharmacological). These models are meant as coarse-grained maps of the brain that translate neuronal architectures (i.e., synaptic connectivity) into parameters, and brain dynamics into belief updating schemes and learning algorithms that update model parameters. Models can be altered in ways that correspond to lesions or interventions and the resultant artificial analogues to behaviour can be safely studied *in silico* (Benrimoh et al., 2019; Parr, Benrimoh, et al., 2018). Of course, the models are inevitably simplified versions of neurobiological systems. When the parameters of these models reproduce psychiatric phenomenology, they constitute computational phenotypes: in other words, they provide analogues of pathological neural phenotypes (Montague et al., 2012). Under active inference, computational phenotypes are generative statistical models that employ Bayesian principles. Crucially, these generative models comprise priors—at many levels—which characterize a particular individual or psychiatric cohort (Adams et al., 2016; Benning, 2015). The models are called *generative* because they generate observable consequences from unobservable causes. On this view, the brain is in the game of inverting or fitting a generative model to her sensory data; namely, inverting the mapping from causes to consequences to infer unobservable states of affairs in the world from their sensed consequences.

Active inference—in theory-driven modelling psychiatry—assumes that the neural processes underlying perception involve inference via the inversion of a generative model (for a discussion of inference under

generative models and classification under *discriminative models*, see: Ng & Jordan, 2002). A generative model is simply the joint probability over the causes and consequences that is usually factorized into a likelihood (i.e., the probability of some sensory consequences, given their causes) and prior beliefs (i.e., the prior probability of some causes or hidden states before seeing sensory data).

Active inference assumes that the brain embodies a generative model of its sensory impressions. Sensory impressions correspond to sensory data (e.g., the activity of wavelength selective photoreceptors), and inference corresponds to the inferred cause of the data (e.g., a color). If priors in the generative model are apt to represent the world, the inference about the causes of the data will provide an accurate account of those data in terms of causes, as simply as possible (technically, with minimal complexity; namely, the difference between prior and posterior beliefs). Suboptimal perceptual inference can arise because of a functionally impaired system (e.g., a lesioned brain), or poorly learned priors (e.g., lack of appropriate training experience or a change in circumstances). Generative models instantiated by the brain are highly complex. They are universally composed of hierarchically organized priors (e.g., they contain priors about low-level causal patterns and higher-level abstractions) that are parameterized to reflect the dynamic structure of the world that they are meant to recapitulate.

Inferring the causes of sensations is but one component of the overall task that the brain has to accomplish. The other key task is to select actions that make inference as efficient as possible. Within the context of modelling brain functions, a generative model will include prior beliefs about transitions between among states of the world (e.g., moving from 'my side of the street' to 'the other side'), given allowable actions (e.g., 'go forward'; 'go backward', etc.). The imperatives for action selection are the same as those for perceptual inference; namely, to maximize the marginal likelihood of sensory data, under the generative model. The only difference is that for policy selection, this likelihood is averaged over the outcomes predicted under the policy in question. The generative model thus can also infer the best course of action, or action policies (i.e., sequences of plausible actions). In short, active inference assumes that, along with many other functions, perception and action are processes of inference in the brain.

An advantage of using computational phenotypes to study psychopathology is that one is forced to give an explicit mathematical description of the dynamics of the pathological functions to phenotype the disorder (Corlett & Fletcher, 2014) (e.g., the neurocognitive process underlying false perceptions, like delusions and hallucinations). Computational phenotyping of this sort simply entails adjusting the priors of the generative model to maximize the likelihood of a particular subject's behaviour or choices (Friston & Penny, 2011; Stephan et al., 2009). Generative models can further simulate psychophysics and neurophysiology (e.g., reaction times and neuromodulatory responses) associated with the hypothesized belief updating mechanisms underlying the pathological function (Parr et al., 2022).

2.3.3 Computational phenotypes beyond the brain

Reflecting the longstanding interest in modelling neural circuitry, active inference is compatible with the aspirations of the RDoC and shares some of the RDoC's assumptions, namely: that pathology can be

understood in terms of circuitry dynamics that adversely affect computational functions, which, typically, subserve adaptive behaviour. Both the RDoC scheme and computational psychiatry borrow from a wealth of experimental work that delineates the different ways in which the brain's processing can go wrong. In recent years, in the hope of adopting a more ecosocial perspective on modelling human cognition, research under active inference has attempted to identify the computational-sociocultural structure of mental disorders. This work has been cashed out in terms of theoretical models and simulation studies of organism-environment interactional behaviour (Bruineberg et al., 2018; Constant, Bervoets, et al., 2018; Constant et al., 2019; Constant, Ramstead, et al., 2018; Friston & Frith, 2015; Kaplan & Friston, 2018; Ramstead et al., 2019; Ramstead et al., 2016). Enlarging the scope of computational phenotyping, these studies have considered the manner in which organisms leverage their environment to support various cognitive functions and forms of social interaction (e.g., communication, social and situated learning, social conformity, cooperative decision-making, joint action, joint attention, etc.)(Veissière et al., 2020).

Conceptually, the ecosocial reading of computational phenotypes is licensed by the fact that the notion of phenotype encompasses levels that reach far beyond the brain (Badcock, 2012; Badcock, Friston, & Ramstead, 2019; Badcock, Friston, Ramstead, et al., 2019; Ramstead et al., 2018, 2019). For instance, a beaver dam is the product of the beaver's behaviour. This behaviour determines the beaver's survival and reproductive success; thereby becoming a target for selective processes. The ensuing combination of agents and their niche is known as an 'extended' phenotype (Dawkins, 1982). An agent can also enter into a coalition (or conflict) with its biotic environment, thereby forming a 'joint' phenotype, wherein no single party owns the phenotype, such as the health state of a parasite host (Queller, 2014), or, presumably, a shared, patterned cultural practice finessed through cultural evolution driving gene-culture co-evolution (Henrich, 2015; Kirmayer et al., 2020a). Ecosocial computational phenotypes rely on such an extended notion of phenotype to model systems beyond individual brains. This, of course, requires translating the ontology of Bayesian neurocomputation (e.g., prior likelihood and inference) to that of human ensembles.

The general idea behind this translation is simple. Just as the brain engages in inference by inverting a generative model of the cause of its sensations, the environment—and the agents it includes—can be regarded as inferring the cause of the sensory impression the environment receives. From the perspective of the environment, the sensory impressions are the agent's actions. Of course, such an anthropomorphic way of talking about the environment is only meant to set up the computational modelling. For instance, a chair may be viewed as providing a series of action possibilities (also known as *affordances* in ecological psychology (Gibson, 1979; Rietveld & Kiverstein, 2014)), each of which yields different agents and context-dependent probabilities. A seat will have a greater probability to elicit the 'sitability' action policy than the 'standability' action in, say, a conference room. In this sense, the chair may be viewed as classifying the action 'sit' under the category, or the cause for the 'agent wanting to sit'. These probabilities are consolidated by histories of agent-environment interactions (e.g., design and construction of the chair, the position of the chair in the room, etc.).

With such a perspective, one can make sense of the many cognitive functions the external world plays for an individual (Constant et al., 2020) and the manner in which typical and atypical cognition may constitutively depend upon those external functions (Constant, Bervoets, et al., 2018). For instance, we know that perceptual

cues guide the acquisition of many cognitive capacities central to normal functioning in social interaction. The production and coordination of perceptual cues such as gestures and uttered narratives guide joint attention during offspring caregiver interaction, and are known to support the acquisition of functions such as folk psychology (which allows a sort of "mind reading" of the states and intentions of others), autobiographical memory, and narrative practices (Fivush & Nelson, 2006; Hutto, 2012; Nelson & Fivush, 2004; Vasil et al., 2020). The failure of the acquisition of such functions is among the popular—although contentious—explanations of the social symptoms of autism (Baron-Cohen et al., 1985; McDonnell et al., 2017). By framing internal and external functions under a single joint phenotype, an ecosocial computational phenotype can explain in a principled fashion (i.e., based on Bayesian principles) the formal relationship between neurocomputational phenomena such as learning and attentional impairments (e.g., Lawson et al., 2014; Van de Cruys et al., 2014), ecological features such as perceptual cues (Constant, Bervoets, et al., 2018), and culturally patterned looping dynamics (Bolis et al., 2017), such as those that characterize interactions between autistic individuals and clinicians or caregivers (Jaswal & Akhtar, 2018).

The inclusion of non-neural factors in theory driven modelling psychiatry allows one to explain the constitutive role of the environment in mind and cognition. This holds the promise of a mechanistic view of mental disorders that can include the computational role of social context and cultural factors. The same approach can be used to model the embodied nature of cognition: namely, the innermost ecology of mind being the brain in the body, which is embedded in the tool-using, interpersonally communicating person that participates in a socially constructed (and populated) niche. The Evolutionary, Cultural, and Computational (ECC) model we consider next brings together the ecosocial reading of computational phenotypes presented above with the adaptationist rationale of evolutionary psychiatry.

3 Evolutionary Computational Ecosocial phenotyping

In section 2, we reviewed the main motivations and principles of evolutionary, cultural, and computational psychiatry. Our aim was to familiarize the reader with the three approaches and their respective modes of explanation and modelling strategies (see summary table 1). In this section, we pursue the integration of cultural and computational psychiatry by supplementing the notion of ecosocial computational phenotypes with an evolutionary interpretation of the structure of generative models. This furnishes an Evolutionary, Cultural, Computational (ECC) model of mental disorders. To illustrate how the ECC model might be applied, we will propose a reading of Major Depressive Disorder (MDD) that articulates the manner in which evolutionary and cultural factors can be integrated into a computational narrative to explain symptoms of MDD.

Before we continue, we should highlight an important distinction. Evolutionary and cultural approaches to psychopathology differ from computational psychiatry in that they both start with assumptions about the potential underlying causes of psychiatric phenomena, while the computational approach can remain agnostic. To the extent that computational modelling is based on a theory of how the brain works (or, at another systemic level,s how social interactions work), it may also make assumptions about causality. However, computational modelling can be used simply to provide a model of observed relationships (i.e.,

input/output mappings) with no presumption that the model describes the actual mechanisms mediating those relationships). When modelling attempts to describe how the brain actually works, the computational model will usually be underwritten by specific evolutionary or socio-cultural accounts of function which have direct consequences for and constraints on the model. The distinction overall is between computational models of the brain as specific ontological theories and computational models as a generic toolkit to capture dynamics that may be instantiated in diverse ways on different substrates. This is precisely one of the ongoing debates around active inference in modelling psychiatry. How much does such modelling assume or entail about the brain as opposed to simply being a flexible framework for (re)describing observed relationships? This is an open debate that we cannot settle here but we note that there are a range of possibilities related to the theoretical or empirical basis for structural relationships that are built into the computational model (rather than those than emerge by virtue of its dynamics).

Put another way, theories about psychiatric phenomena in evolutionary and cultural psychiatry aim to account for the specific aetiological, phenomenological, and nosological relations between observed symptoms and their underlying causes—and associated syndromes—by drawing from specific accounts of human (pre)history, development, and current social contexts, whereas computational psychiatry can be used to model symptoms by incorporating a variety of possible underlying causes. In that sense, computational psychiatry constitutes a flexible method to analyse psychiatric disorders, rather than a substantive theory of their ontology or aetiology.

The upshot of this is that computational psychiatry, or computational phenotyping under active inference, furnishes a way to integrate, within a single coherent and principled framework, a variety of theories about psychiatric phenomena. The ECC approach should not be viewed as a single implementable computational model, but rather, as a description of the variety of priors one could use to parameterize computational phenotypes that conform to the principles of evolutionary and cultural psychiatry. For a simulation study of such a computational phenotype see (Constant et al., 2021).

Table 1. Modes of Explanation and Modelling Strategies

Discipline and Mode of Explanation	**Focus with respect to Wakefield's definition**	**Conception of mental disorders**	**Modelling strategy**
Evolutionary psychiatry (selected account of function and Darwinian rationales)	The concept of dysfunction	Developmentally aggravated vulnerabilities understood as proximate causes shaped by ultimate causes	Darwinian rationales (cf. Box 1)
Cultural psychiatry (looping effects of human kinds, impact of self-construal, and	The concept of the harmful	Behavioural patterns causing psychological distress and functional impairment configured at the subjective level, and shaped by	Ecosocial model

ecosocial systemic models)		socionormative interactions, and cultural affordances	
Computational psychiatry (active inference and ecosocial phenotyping in theory drive modelling psychiatry)	The concept of dysfunction; potential to model how harm and dysfunction interact	Suboptimal inference of perception and action caused by lesion or atypically learned model paramters	Computational phenotyping

3.1 Evolution and culture in ECC

3.1.1 Evolution in ECC

The ecosocial reading of computational phenotypes can be supplemented with an evolutionary interpretation. Internal priors of a generative model can be viewed as targets for selection; they can be studied as (epi)genetic, structural, or *adaptive priors* (Friston, 2010; Friston & Stephan, 2007). Adaptive priors are endowed by evolution and have been geared towards adapting the individual to the ancestral environment. They can be contrasted with (empirical) priors which are learned over developmental time via experience-dependent neuronal plasticity.

From a modelling perspective, the consequence of this is that adaptive priors will exert a strong top-down influence over empirical priors that can be learned, and thus over behaviour and neurophysiology. For instance, our prior preferences for energy-rich food can be viewed as an innate prior that will be paired with the learned empirical prior beliefs about the probability of finding energetic resources in the current environment (Richerson, 2018). Such an adaptationist rationale as applied to priors is useful for designing pathological generative models under the views of mismatch theory, constraints and trade-offs argued by evolutionary psychiatry.

3.1.2 Culture in ECC

With respect to culture, we have seen that those states external to the generative model representing the environment, can be modelled in terms of priors and likelihoods, and thus the environment could itself be read as learning about its denizens. This view underwrites the ecosocial interpretation of computational phenotypes. Culture is defined as shared knowledge, practices, values, and institutions that constitute the way of life of a group of individuals or community (Kirmayer, 2018). From a computational perspective, culture may thus be modelled as the calibration (viz. practice) between the priors, likelihood, and agents constituting the environment (viz. institutions) and the priors, likelihood and sensations making up the agents themselves (viz. knowledge and values)(Constant et al., 2019; Ramstead et al., 2016; Roepstorff et al., 2010; Veissière, 2016; Veissière & Stendel, 2018; Veissière et al., 2020).

The calibration of generative models is mediated by the exchange of sensory cues generated by the environment and actions generated by the agent. Over time, this exchange should attune the generative model of the agent to her environment (Constant, Ramstead, et al., 2018; Ramstead et al., 2019; Veissière et al., 2020). Cultural models such as those construed by cultural psychiatry (i.e., the stable discursive and expressive styles of illness experience encoded in cognitive schema, practices, and social institutions) may thus be viewed as illness-specific calibrations of agents and world's generative models, which consolidate through ecosocial looping dynamics.

3.1.3 Mechanism and function in ECC

The ECC considers model parameters that reflect biological (and cultural) phenomena caused by proximal factors (e.g., mechanisms) and ultimate factors (e.g., functions). This distinction between proximate and ultimate factors, as discussed earlier, is one of the ways in which evolutionary psychiatry tries to understand "why evolution left us with traits that make us vulnerable to mental disorders." The taxonomy of priors described in this section—i.e., adaptive priors, vs. empirical and environmental or cultural ones—could be misconstrued as promoting a false dichotomy between proximate and ultimate causes: adaptive priors are meant to reflect the species' evolutionary history (its phylogeny), while empirical priors are meant to reflect the way an organism learns its environment over development (its ontogeny). This way of thinking is problematic, however, because it suggests an overly simplistic way to think about adaptation and development.

In particular, the notion of adaptive priors used here might be misread as meaning an "innate" prior, a consequence of our evolutionary history, which is a controversial notion that certainly cannot cover many of the kinds of priors relevant to psychiatric disorders. In our model, adaptive priors are distinguished from purely learned, empirical or developmental priors. Historically, the folk concept of innateness has often conflated notions that reflect distinct and often irreconcilable biological realities (Griffiths, 2002). Those notions include (i) developmental fixity (i.e., the idea that an innate trait is 'hard to change'); (ii) species nature (i.e., the idea that an innate trait is 'universal'), and (iii) intended outcome (i.e., the idea that an innate trait is 'there by design'). Appealing, either implicitly or explicitly, to such a folk essentialist way of thinking in science runs the risk of unjustifiably importing conclusions based on findings in one domain of biology into another disjoint domain (e.g., "because this trait is universal, it must be there by design, and because it is there by design, it will not change over development")(Griffiths, 2002). It is precisely these risks that the kind of computational phenotyping proposed here contends with, as it integrates model parameters that are meant to reflect 'adaptive' versus 'learnable' traits.

The ECC, however, circumvents the problem of folk essentialism because the notion of an adaptive prior simply refers to a temporal scale of organization relative to a scale of interest. An adaptive prior is one that performs an evolutionary function (for a review of the notion of function, see: Christie et al., 2021) and for that reason, it is reliably transmitted to individuals from one generation to the next (e.g., the hierarchical structure and plasticity of the developing brain)(Badcock, Friston, & Ramstead, 2019). By contrast, an empirical prior is

limited to (or learned during) the life span of the system of interest (e.g., a given connection pattern among neurons), and may not be passed on to subsequent generations. Of course, this implies that social systems or niches and cultural contexts that may have temporal duration beyond the life of an individual—and that are passed on exogenetically to the next generation—may also contribute scales of organization relevant to explaining psychopathology (Kirmayer et al., 2020a).

Thus, adaptive priors are typically 'hard to change' (that is, appear to be developmentally fixed) may simply be 'slow to change'; hence, developmental fixity does not suppose a "species' nature", as that trait may change over phylogenetic time. Universality just refers to the fact that the adaptive prior will be spread across a population for a period extending beyond the individual life span of the members of that population. It denotes the phenotypic synchrony among individuals sharing the adaptive prior within a given (intergenerational) timeframe. Finally, the notion of 'design' refers to the evolutionary function of the trait and is manifested by the top-down influence that the adaptive prior will exert on empirical priors (e.g., computationally, for one update at the adaptive level, there might be multiple updates at the empirical level).

At this juncture, it is worth noting that the ECC approach outlined here appeals to a multiscale model of the human brain, called 'the hierarchically mechanistic mind', which explains cognition and behaviour by integrating active inference with Tinbergen's four questions in biology (i.e., adaptation, phylogeny, ontogeny, and mechanisms)(Badcock, Friston, & Ramstead, 2019; Badcock, Friston, Ramstead, et al., 2019). According to this perspective, understanding the computational processes that underlie human action and perception requires an integrative approach that captures the evolutionary, developmental, and real-time dynamics that govern them. By incorporating both adaptive and empirical priors in a single modelling approach, the ECC presents an empirically viable avenue to help researchers unpack the complexities of Tinbergen's four questions. We suggest, therefore, that our modelling approach might not only be of interest to researchers in psychiatry, but also to those in the human and biological sciences more broadly.

3.2 Major Depressive Disorder under the ECC

Common targets of computational phenotyping include schizophrenia (Benrimoh & Friston, 2020), autism (Friston, 2017) and Major depressive disorder (MDD) (Huys, Daw, et al., 2015; Huys, Guitart-Masip, et al., 2015). Evolutionary (Darwinian) and cultural mechanistic explanations have already been proposed to account for the symptoms and syndrome of depression (Allen & Badcock, 2006; Badcock et al., 2017).

Computational psychiatry models the core symptoms of MDD (e.g., diminished drive, loss of energy, anhedonia) in terms of computational failings in the evaluation of long-term utility reward functions, a.k.a. the evaluation of *secondary utility* (Huys, Daw, et al., 2015). Secondary utility relates to the value of stimuli whose reward causal structure is complex and spatiotemporally extended (e.g., the reward value of accumulating money). On the other hand, primary, biological, or 'hedonic' utility –as opposed to secondary, 'anticipatory' utility –relates to reward that is a proxy for reproductive success and survival (e.g., avoiding pain; seeking energy rich food) (Huys, Daw, et al., 2015), thereby relating to adaptive priors and preferences that (under adaptationist assumptions) have increased reproductive success in the past. This is consistent with

evolutionary approaches to mood disorder arguing for the adaptive value of low mood rather than MDD *per se* (Allen & Badcock, 2003, 2006; Badcock et al., 2017).

3.2.1 Pessimistic priors

One computational pathway to understanding MDD as a dysfunction of long-term reward evaluation is the acquisition of pessimistic priors that entail biased learning of environmental states. The main function of priors—in a generative model—is to disambiguate the sensory information the system receives, in order to perform successful inference and select adaptive action. For instance, as per our problem of indirect perception, one cannot directly infer the mood of another person solely from the sensory information that person's face affords. Rather, one must take into account some high-level assumptions about the person's behaviour over time (e.g., "she is usually a smiling person, but now her smile must mean something different because of what I said yesterday").

In other words, priors always bias the way we treat incoming information, and consequently, the way one selects action towards future sampling of the environment (e.g., "perhaps I should avoid talking to her as I'm sure she will reject me"). In MDD, priors biasing such model-based decision-making are priors that tip the balance towards pessimistic inference, thereby leading to systematic pessimistic thoughts (a.k.a., a negative thinking bias,(Teasdale, 1983). For instance, MDD patients form negative sentences more frequently and faster than healthy controls, when presented with optimistic and pessimistic options (e.g., in the scramble sentence test)(Hindash & Amir, 2012; Rude et al., 2003). As we will see next, pessimistic thoughts may interact with depressive rumination, and lead to the downward depressive spiral of negative expectations and self-evaluation, anhedonia, social withdrawal, and the suppression of reward-approach behaviour characteristic of MDD. This is explained in terms of the autodidactic installation of pessimistic priors.

3.2.2 Reinforcing pessimistic priors

Many symptoms of depression are commonly experienced by healthy individuals and become a target for psychiatric MDD diagnosis only when they become enduring and lead to clinically significant functional impairment. Therefore, any account of depression should explain the maintenance of MDD symptoms over time. Another role of priors is to guide attention towards sensory cues deemed informative, given these same priors (Feldman & Friston, 2010), a.k.a., *self-evidencing* (Hohwy, 2016). Explicitly engaged, or endogenous attention, for instance, can be viewed as a form of internal action (Brown et al., 2013; Edwards et al., 2012; Limanowski & Friston, 2018; Parr & Friston, 2019) that assesses the relevance of information, sometimes in a biased fashion (Hohwy, 2013; Paton et al., 2012; Hohwy, 2013; Paton et al., 2012). In MDD patients, aversive events invoke more recurrent and persistent cognitive processing. For instance, depressed patients gaze longer at negative stimuli: i.e., stimuli or information about negative outcomes (Caseras et al., 2007) and spend more time examining them (Kellough et al., 2008). They also report less positive emotion in response to positive images and more arousal to aversive images (Sloan et al., 1997). Sustained endogenous attention over

negative stimuli suggests that aversive events are considered informative, that is, disambiguating with respect to pessimistic priors (Huys, Daw, et al., 2015).

Recurrent sampling of negative information necessarily entails reduced sampling of positive information (Huys, Daw, et al., 2015); the sampling of information being one of the two ways in which one learns and update priors—our bias that drives appraisal of the world (the other being the pruning, or synaptic homeostasis, that underlies structure learning (see: Friston, Lin, et al., 2017 and Tononi & Cirelli, 2006). Ongoing learning based on negative information is characteristic of the inability to inhibit rumination, defined as the tendency to focus on one's depressive state, along with the causes, meanings, and consequences of one's depression (Nolen-Hoeksema, 1991). Interestingly, rumination is often motivated by the belief that ruminating will bring insights into how to solve the cause of rumination (Lyubomirsky & Nolen-Hoeksema, 1993).

The maintenance of MDD symptoms may be explained by the looping effect that underlies the autodidactic learning of pessimistic priors, when considered from the point of view of the computational machinery of the brain embedded in the social world. The loop is simple: pessimistic priors bias attention and learning, which biases active sampling towards rumination and exogenous negative information that confirm the pessimistic prior (i.e., self-evidences it), thereby leading to the consolidation of this pessimistic prior over time (i.e., minimization of uncertainty based on information that confirms the prior) (Huys, Daw, et al., 2015). Exogenous negative information propagates in the social world through public discourse as depression becomes an increasingly popular diagnostic label, and characteristic idioms of distress are used by sufferers to frame their experience and guide their attention towards that which conforms to these idioms (Kirmayer et al., 2017). In so doing, institutionally sanctioned negative exogenous information shapes the way one attends to one's own experience, body, and sensations, thereby reinforcing those priors' beliefs about one's illness.

Indeed, depressive patients are able to leverage and apply emotion regulation strategies to tackle their affliction when they are instructed to do so, but have difficulties selecting such strategies on their own (Ehring et al., 2010). This speaks to the role of the social environment in the maintenance of MDD. It further speaks to the need to model computational looping effects of depression under ECC, not only in terms of learning and action selection dynamics in the generative model, but also in terms of environmental dynamics that feed back into learning to influence subsequent action selection.

3.2.3 Pessimistic priors and adaptive priors

We have seen that the general mechanism that underwrites MDD may be the maintenance and reinforcement of a pessimistic prior. From a behavioural point of view, the sampling of negative information and rumination reinforces the pessimistic prior. In return, the pessimistic prior further orients the person towards actions that will sample negative information, which accounts for the downward spiral characteristic of MDD. From a cultural point of view, the spiral may be consolidated through pathological cognition. This process is driven by endogenous and exogenous attention: because of pessimistic beliefs, the person attends to negative stimuli, and in return, negative stimuli that confirms the pessimistic beliefs become increasingly available in her environment, social niche or cultural context. In effect, the diagnostic category becomes an organizing

framework for experience that exerts its own effects in the cycles that constitute depressive cognition (Kirmayer et al., 2017)[11]. Of course, this is not the only (or main effect) of culture, which also creates social-structural conditions of adversity and modes of adaptation that engender the vicious cycles of depression (Kirmayer & Jarvis, 2007).

From an evolutionary point of view, given the survival value of being able to rapidly attend to potentially threatening information (Dijksterhuis & Aarts, 2003), the learning of a pessimistic prior can be further precipitated by a predisposition to seek negative stimuli, or evidence that will confirm the source of such pessimism. This predisposition can be modelled as a prior preference for the source of negative stimuli. This was demonstrated by Constant and colleagues (2021) in a computational study of the pathogenesis of MDD. They simulated a 'social' two-armed bandit scenario, in which the player had to decide which of two social partners to visit. Each partner afforded a level of reward from low to high, and an associated level of uncertainty over whether the visit would afford a high or a low reward. This setting was meant to reflect uncertainty in environmental contingencies, corresponding to the changing mood of social partners. At the outset, the synthetic agent performed the task adaptively and learned optimistic beliefs, until an adverse life event—that increased social volatility—perturbed social contingencies. Learned optimistic beliefs then shifted to pessimistic beliefs, as the agent kept receiving low reward when approaching social partners believed to afford high reward. As the simulation unfolded, expected utility went down, and eventually, the agents stopped engaging altogether, thereby evincing severe social withdrawal and low expected utility characteristic of MDD. Crucially, to reach the MDD state, the agent had to be endowed with a fixed prior preference for high social reward that would incentivize her to keep exposing herself to social partners, despite continued negative evidence (or outcomes). From an ECC point of view, the fixed prior preference played the role of an adaptive prior, which, under normal circumstances, fosters social interactions. However, under abnormal circumstances, for instance, when social volatility increases and persists, the same adaptive prior will generate behaviour that engenders low mood and eventually MDD. Accordingly, the pathogenesis of MDD in Constant et al. (2021) could be read under the mismatch rationale discussed above. Importantly, this computational study exemplifies our ECC approach by showing how evolutionary and empirical priors that reflect current social-cultural contexts can interact to produce generative models, characteristic of psychiatric disorder.

4 Concluding remarks: towards an integrative systemic view of mental disorder

In this chapter, we have entertained a trialogue between three approaches to psychiatry: evolutionary, cultural, and computational. We have focused on themes central to these approaches, such as adaptationist thinking, looping effects, and generative models in computational phenotyping. We have suggested a way to merge these perspectives under an Evolutionary Cultural Computational (ECC) model that characterizes the extended phenotype of the individual in context. The goal of this exercise was to exemplify an ecosocial

[11] This kind of looping effect provides a key illustration of how a social-cultural perspective enriches the evolutionary computational model.

computational model of mental disorders that harmonizes the constructs of evolutionary, cultural, and computational psychiatry, integrating their respective systemic views into a systemic model.

While we believe the ECC approach provides a framework for integrating diverse perspectives in psychiatric theory and research, it has a number of important limitations. The ECC approach puts few constraints on theory building and an ECC computational model will only be as accurate as the evolutionary and cultural models that inform it. Computational models are technically challenging and require specific training to conduct analyses, which not be part of the background of those with the requisite expertise in evolutionary or cultural psychiatry. As a method of building hypothetical models, the validity of ECC cannot be directly tested. Ultimately, its validity rests on its scientific and practical utility for generating new models. That said, the performance of any one ECC model can be compared again competing models and real-world data to confirm or refute simulation outcomes. Translating computational models to psychiatric practice presents its own challenges, which might be met by developing diagnostic and assessment tools that allow practitioners to use client data to predict the course of illness in different social contexts or under different treatment conditions.

Despite those limitations, we believe that there are several ways in which the proposed ECC model can contribute to psychiatric theory, research, and practice. active inference models in computational psychiatry are meant to function as heuristic descriptions of the brain. Based on these heuristics, one can simulate pathological behaviour and test, *in silico*, various interventions that mimic the effects of pharmacological agents, psychotherapy, social interventions, or other treatments on model parameters to examine the potential efficacy of this treatment to return the agent to 'normal' functioning. Such modelling can suggest the sensitivity of illness trajectories to particular types of intervention and the potential interactions among multiple interventions.

Because the ECC model considers evolutionary and cultural parameters, *in silico* testing of an ECC model may provide new insights into the potential efficacy of interventions in more ecologically valid contexts (e.g., for a simulation study applying an ECC model to depression, see: Constant et al., 2021). ECC phenotyping methods can be used to simulate specific kinds of suboptimal perceptual inference (e.g., the misinterpretation of a social partner's intention) that may be associated with psychiatric disorders by considering the influence of parameters reflecting the neural, developmental, evolutionary, and social dimensions of a phenotype. ECC phenotyping methods can also be used to identify clinically relevant phenotypes by fitting simulations to large datasets harvested from a range of different contexts, including: data drawn from interactions in shared environments such as social media platforms (which would reflect the manner in which people engage in a shared generative process); data drawn from psychophysics (e.g., eye tracking and response time data); and imaging or EEG data (which would reflect the impact of individuals' generative models on behaviour). Using standard methods for Bayesian model comparison (e.g., Bayesian model reduction (Friston et al., 2016, 2018), researchers could compare ECC phenotypes in terms of their model evidence, each emphasizing different components of the phenotype.

Finally, returning to the problem of disciplinary boundaries discussed at the outset, the ECC model—understood as a multidisciplinary platform to integrate diverse approaches to psychiatric phenomena

through the same computational model—could allow practitioners with various backgrounds to see how their perspectives can connect and converge; thereby enriching each other's ways of thinking about psychiatric disorders. Indeed, the goal of the ECC is to allow researchers and clinicians consider how phenomena like adaptation can contribute conceptually to an understanding of culture, and *vice versa,* that is, how cultural context and meaning shape the exigencies and outcomes of adaptation in health and illness. The human mind is a complex structure with nested levels of organization and boundaries that reflect our cultural co-evolution and varied forms of social life. If we are to come to grips with the difficulties in adaptation and functioning that are the domain of psychiatry, we must develop tools that capture such complexities.

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, and Psychiatry*, *87*(1), 53–63.

Adriaens, P. R., & De Block, A. (2010). The evolutionary turn in psychiatry: a historical overview. *History of Psychiatry*, *21*(82 Pt 2), 131–143.

Alegría, M., NeMoyer, A., Falgàs Bagué, I., Wang, Y., & Alvarez, K. (2018). Social Determinants of Mental Health: Where We Are and Where We Need to Go. *Current Psychiatry Reports*, *20*(11), 95.

Allen, N. B., & Badcock, P. B. T. (2003). The social risk hypothesis of depressed mood: evolutionary, psychosocial, and neurobiological perspectives. *Psychological Bulletin*, *129*(6), 887–913.

Allen, N. B., & Badcock, P. B. T. (2006). Darwinian models of depression: a review of evolutionary accounts of mood and mood disorders. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, *30*(5), 815–826.

Badcock, P. B. (2012). Evolutionary systems theory: A unifying meta-theory of psychological science. *Review of General Psychology: Journal of Division 1, of the American Psychological Association*, *16*(1), 10–23.

Badcock, P. B., Davey, C. G., Whittle, S., Allen, N. B., & Friston, K. J. (2017). The Depressed Brain: An Evolutionary Systems Theory. *Trends in Cognitive Sciences*, *21*(3), 182–194.

Badcock, P. B., Friston, K. J., & Ramstead, M. J. D. (2019). The hierarchically mechanistic mind: A free-energy formulation of the human psyche. *Physics of Life Reviews*, *31*, 104–121.

Badcock, P. B., Friston, K. J., Ramstead, M. J. D., Ploeger, A., & Hohwy, J. (2019). The hierarchically mechanistic mind: an evolutionary systems theory of the human brain, cognition, and behaviour. *Cognitive, Affective & Behavioral Neuroscience*. https://doi.org/10.3758/s13415-019-00721-3

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a “theory of mind” ? *Cognition*, *21*(1), 37–46.

Bateson, P. (2001). Fetal experience and good adult design. *International Journal of Epidemiology*, *30*(5), 928–934.

Bateson, P., & Laland, K. N. (2013). Tinbergen’s four questions: an appreciation and an update. *Trends in Ecology & Evolution*, *28*(12), 712–718.

Bechtel, W. (2009). Explanation: Mechanism, Modularity, and Situated Cognition. In M. Aydede & P. Robbins (Eds.), *The Cambridge Handbook of Situated Cognition* (pp. 155–170). Cambridge: Cambridge University Press.

Benning, T. B. (2015). Limitations of the biopsychosocial model in psychiatry. *Advances in Medical Education and Practice*, *6*, 347–352.

Benrimoh, D. A., & Friston, K. J. (2020). All grown up: Computational theories of psychosis, complexity, and progress. *Journal of Abnormal Psychology*, *129*(6), 624.

Benrimoh, D., Parr, T., Adams, R. A., & Friston, K. (2019). Hallucinations both in and out of context: An active inference account. *PloS One*, *14*(8), e0212379.

Bolis, D., Balsters, J., Wenderoth, N., Becchio, C., & Schilbach, L. (2017). Beyond Autism: Introducing the Dialectical Misattunement Hypothesis and a Bayesian Account of Intersubjectivity. *Psychopathology*, *50*(6). https://doi.org/10.1159/000484353

Bolton, D., & Gillett, G. (2019). *The Biopsychosocial Model of Health and Disease: New Philosophical and Scientific Developments*. Springer International Publishing.

Boorse, C. (1982). On the Distinction between Disease and Illness. In *Medicine and Moral Philosophy* (pp. 3–22). https://doi.org/10.1515/9781400853564.3

Borsboom, D., Cramer, A. O. J., & Kalis, A. (2019). Brain disorders? Not really: Why network structures block reductionism in psychopathology research. *The Behavioural and Brain Sciences*, *42*. https://doi.org/10.1017/S0140525X17002266

Bourrat, P., & Griffiths, P. E. (2021). *The idea of mismatch in evolutionary medicine*. http://philsci-archive.pitt.edu/18933/

Brown, H., Adams, R. A., Parees, I., Edwards, M., & Friston, K. J. (2013). Active inference, sensory attenuation and illusions. *Cognitive Processing*, *14*(4), 411–427.

Bruineberg, J., Rietveld, E., Parr, T., van Maanen, L., & Friston, K. J. (2018). Free-energy minimization in joint agent-environment systems: a niche construction perspective. *Journal of Theoretical Biology*. https://doi.org/10.1016/j.jtbi.2018.07.002

Canter, D. (2012). Challenging neuroscience and evolutionary explanations of social and psychological processes. *Contemporary Social Science*, *7*(2), 95–115.

Caseras, X., Garner, M., Bradley, B. P., & Mogg, K. (2007). Biases in visual orienting to negative and positive scenes in dysphoria: An eye movement study. *Journal of Abnormal Psychology*, *116*(3), 491–497.

Christie, J. R., Brusse, C., Bourrat, P., Takacs, P., & Griffiths, P. E. (2021). *Are biological traits explained by their "selected effect" functions?* http://philsci-archive.pitt.edu/19832/

Compton, W. M., & Guze, S. B. (1995). The neo-Kraepelinian revolution in psychiatric diagnosis. In *European Archives of Psychiatry and Clinical Neuroscience* (Vol. 245, Issues 4-5, pp. 196–201). https://doi.org/10.1007/bf02191797

Constant, A., Bervoets, J., Hens, K., & Van de Cruys, S. (2018). Precise Worlds for Certain Minds: An Ecological Perspective on the Relational Self in Autism. *Topoi. An International Review of Philosophy*. https://doi.org/10.1007/s11245-018-9546-4

Constant, A., Clark, A., Kirchhoff, M., & Friston, K. J. (2020). Extended active inference: Constructing predictive cognition beyond skulls. *Mind & Language*, *mila.12330*. https://doi.org/10.1111/mila.12330

Constant, A., Hesp, C., Davey, C. G., Friston, K. J., & Badcock, P. B. (2021). Why Depressed Mood is Adaptive: A Numerical Proof of Principle for an Evolutionary Systems Theory of Depression. *Computational Psychiatry (Cambridge, Mass.)*, *5*(1), 60–80.

Constant, A., Ramstead, M. J. D., Veissière, S., & Friston, K. J. (2019). Regimes of Expectations: An active inference Model of Social Conformity and Decision Making. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2019.00679

Constant, A., Ramstead, M. J. D., Veissière, S. P. L., Campbell, J. O., & Friston, K. J. (2018). A variational approach to niche construction. *Journal of the Royal Society, Interface / the Royal Society*, *15*(141). https://doi.org/10.1098/rsif.2017.0685

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.

Craver, C. F. (2006). When mechanistic models explain. *Synthese*, *153*(3), 355–376.

Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., & Friston, K. (2020). active inference on discrete state-spaces: A synthesis. *Journal of Mathematical Psychology*, *99*, 102447.

Dawkins, R. (1982). *The extended phenotype*. Oxford University Press.

Del Giudice, M. (2014). An Evolutionary Life History Framework for Psychopathology. *Psychological Inquiry*, *25*(3-4), 261–300.

Dijksterhuis, A., & Aarts, H. (2003). On Wildebeests and Humans: The Preferential Detection of Negative Stimuli. *Psychological Science*, *14*(1), 14–18.

Durisko, Z., Mulsant, B. H., & Andrews, P. W. (2015). An adaptationist perspective on the aetiology of depression. *Journal of Affective Disorders*, *172*, 315–323.

Eckersley, R. (2006). Is modern Western culture a health hazard? In *International Journal of Epidemiology* (Vol. 35, Issue 2, pp. 252–258). https://doi.org/10.1093/ije/dyi235

Eckersley, R. (2011). Troubled youth: an island of misery in an ocean of happiness, or the tip of an iceberg of suffering? In *Early Intervention in Psychiatry* (Vol. 5, pp. 6–11). https://doi.org/10.1111/j.1751-7893.2010.00233.x

Edwards, M. J., Adams, R. A., Brown, H., Pareés, I., & Friston, K. J. (2012). A Bayesian account of "hysteria." *Brain: A Journal of Neurology*, *135*(Pt 11), 3495–3512.

Ehring, T., Tuschen-Caffier, B., Schnülle, J., Fischer, S., & Gross, J. J. (2010). Emotion regulation and vulnerability to depression: spontaneous versus instructed use of emotion suppression and reappraisal. *Emotion* , *10*(4), 563–572.

Engel, G. L. (1981). The Clinical Application of the Biopsychosocial Model. In *Journal of Medicine and Philosophy* (Vol. 6, Issue 2, pp. 101–124). https://doi.org/10.1093/jmp/6.2.101

Faucher, L., & Forest, D. (2021). *Defining Mental Disorder: Jerome Wakefield and His Critics*. MIT Press.

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free energy. *Frontiers in Human Neuroscience*, *4*, 215.

First, M. B. (2012). The development of DSM-III from a historical/conceptual perspective. In *Philosophical Issues in Psychiatry II* (pp. 127–140). https://doi.org/10.1093/med/9780199642205.003.0020

Fivush, R., & Nelson, K. (2006). Parent–child reminiscing locates the self in the past. *The British Journal of Developmental Psychology*, *24*(1), 235–251.

Fleck, L. (1979). *Genesis and Development of a Scientific Fact*. University of Chicago Press.

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nature Reviews. Neuroscience*, *11*(2), 127–138.

Friston, K. J. (2017). Precision Psychiatry [Review of *Precision Psychiatry*]. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging*, *2*(8), 640–643.

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: a process theory. *Neural Computation*, *29*(1), 1–49.

Friston, K. J., & Frith, C. D. (2015). Active inference, communication and hermeneutics. *Cortex; a Journal Devoted to the Study of the Nervous System and Behaviour*, *68*, 129–143.

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., & Ondobaka, S. (2017). Active Inference, Curiosity and Insight. *Neural Computation*, *29*(10), 2633–2683.

Friston, K. J., Litvak, V., Oswal, A., Razi, A., Stephan, K. E., van Wijk, B. C. M., Ziegler, G., & Zeidman, P. (2016). Bayesian model reduction and empirical Bayes for group (DCM) studies. *NeuroImage*, *128*, 413–431.

Friston, K. J., Parr, T., & Zeidman, P. (2018). Bayesian model reduction. In *arXiv [stat.ME]*. arXiv. http://arxiv.org/abs/1805.07092

Friston, K. J., & Penny, W. (2011). Post hoc Bayesian model selection. *NeuroImage*, *56*(4), 2089–2099.

Friston, K. J., & Stephan, K. E. (2007). Free-energy and the brain. *Synthese*, *159*(3), 417–458.

Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014). Computational psychiatry: the brain as a phantastic organ. *The Lancet. Psychiatry*, *1*(2), 148–158.

Gauld, C., Dumas, G., Fakra, É., Mattout, J., & Micoulaud-Franchi, J.-A. (2021). Les trois cultures de la psychiatrie computationnelle. *Annales Médico-Psychologiques, Revue Psychiatrique*, *179*(1), 63–71.

Ghaemi, S. N. (2009). The rise and fall of the biopsychosocial model. British Journal of Psychiatry, *195*(1), 3–4. https://doi.org/10.1192/bjp.bp.109.063859

Ghaemi, S. N. (2010). *The rise and fall of the biopsychosocial model: reconciling art and science in psychiatry*. JHU Press.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin.

Glennan, S., Illari, P., & Weber, E. (2021). Six Theses on Mechanisms and Mechanistic Science. *Journal for General Philosophy of Science. Zeitschrift Fur Allgemeine Wissenschaftstheorie*. https://doi.org/10.1007/s10838-021-09587-x

Gould, S. J. (1991). Exaptation: A crucial tool for evolutionary analysis. *The Journal of Social Issues*, *47*(3), 43–65.

Gould, S. J., & Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character. Royal Society* , *205*(1161), 581–598.

Graves, J. L., Jr, Reiber, C., Thanukos, A., Hurtado, M., & Wolpaw, T. (2016). Evolutionary Science as a Method to Facilitate Higher Level Thinking and Reasoning in Medical Training. *Evolution, Medicine, and Public Health*. https://doi.org/10.1093/emph/eow029

Griffiths, P. E. (2002). What is innateness? *The Monist*, *85*(1), 70–85.

Grunspan, D. Z., Nesse, R. M., Barnes, M. E., & Brownell, S. E. (2018). Core principles of evolutionary medicine: A Delphi study [Review of *Core principles of evolutionary medicine: A Delphi study*]. *Evolution, Medicine, and Public Health*, *2018*(1), 13–23.

Guillemin, M., & Barnard, E. (2015). George Libman Engel: The Biopsychosocial Model and the Construction of Medical Practice. In *The Palgrave Handbook of Social Theory in Health, Illness and Medicine* (pp. 236–250). https://doi.org/10.1057/9781137355621_15

Hacking, I. (1985). Making Up People. In M. S. A. D. E. W. Heller (Ed.), *Reconstructing Individualism.* Stanford University Press.

Hacking, I. (1995). The looping effect of human kinds. In D. Sperber, D. Premack, & A. J. Premack (Ed.), *Causal cognition: A multidisciplinary debate* (pp. 351–383). Oxford University Press.

Hacking, I. (2000). *The social construction of what?* Harvard University Press.

Henrich, J. (2015). *The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

Henriques, G. (2011). *A new unified theory of psychology*. Springer.

Hindash, A. H. C., & Amir, N. (2012). Negative Interpretation Bias in Individuals with Depressive Symptoms. *Cognitive Therapy and Research*, *36*(5), 502–511.

Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

Hohwy, J. (2016). The self-evidencing brain. *Noûs*, *50*(2), 259–285.

Houston, A. I., McNamara, J. M., & I., H. A. (1999). *Models of Adaptive Behaviour: An Approach Based on State*. Cambridge University Press.

Hutchins, E. (2010). Cognitive ecology. *Topics in Cognitive Science*, *2*(4), 705–715.

Hutto, D. D. (2012). *Folk psychological narratives: The sociocultural basis of understanding reasons*. MIT Press.

Huys, Q. J. M., Daw, N. D., & Dayan, P. (2015). Depression: a decision-theoretic analysis. *Annual Review of Neuroscience*, *38*, 1–23.

Huys, Q. J. M., Guitart-Masip, M., Dolan, R. J., & Dayan, P. (2015). Decision-Theoretic Psychiatry. *Clinical Psychological Science*, *3*(3), 400–421.

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.

Insel, T. R. (2014). The NIMH Research Domain Criteria (RDoC) Project: Precision Medicine for Psychiatry. In *American Journal of Psychiatry* (Vol. 171, Issue 4, pp. 395–397). https://doi.org/10.1176/appi.ajp.2014.14020138

Jaswal, V. K., & Akhtar, N. (2018). Being vs. Appearing Socially Uninterested: Challenging Assumptions about Social Motivation in Autism. *The Behavioural and Brain Sciences*, 1–84.

Kaplan, R., & Friston, K. J. (2018). Planning and navigation as active inference. *Biological Cybernetics*. https://doi.org/10.1007/s00422-018-0753-2

Karasewich, T. A., & Kuhlmeier, V. A. (2020). Trait social anxiety as a conditional adaptation: A developmental and evolutionary framework. In *Developmental Review* (Vol. 55, p. 100886). https://doi.org/10.1016/j.dr.2019.100886

Keller, M. C., & Miller, G. (2006). Resolving the paradox of common, harmful, heritable mental disorders: which evolutionary genetic models work best? *The Behavioural and Brain Sciences*, *29*(4), 385–404; discussion 405–452.

Kellough, J. L., Beevers, C. G., Ellis, A. J., & Wells, T. T. (2008). Time course of selective attention in clinically depressed young adults: an eye tracking study. *Behaviour Research and Therapy*, *46*(11), 1238–1243.

Kendler, K. S., Parnas, J., & Zachar, P. (2020). *Levels of Analysis in Psychopathology: Cross-Disciplinary Perspectives*. Cambridge University Press.

Kenrick, D. T. (2001). Evolutionary Psychology, Cognitive Science, and Dynamical Systems: Building an Integrative Paradigm. *Current Directions in Psychological Science*, *10*(1), 13–17.

Kirmayer, L. J. (2005). Culture, context and experience in psychiatric diagnosis. *Psychopathology*, *38*(4), 192–196.

Kirmayer, L. J. (2006). Beyond the "New Cross-cultural Psychiatry": Cultural Biology, Discursive Psychology and the Ironies of Globalization. In *Transcultural Psychiatry* (Vol. 43, Issue 1, pp. 126–144). https://doi.org/10.1177/1363461506061761

Kirmayer, L. J. (2007). Psychotherapy and the Cultural Concept of the Person. In *Transcultural Psychiatry* (Vol. 44, Issue 2, pp. 232–257). https://doi.org/10.1177/1363461506070794

Kirmayer, L. J. (2015). Re-visioning psychiatry: Towards an ecology of mind in health and illness. In L. J. Kirmayer, R. Lemelson, & C. A. Cummings (Eds.), *Cultural Phenomenology, Critical Neuroscience and Global Mental Health* (pp. 622–660). Cambridge University Press.

Kirmayer, L. J. (2018). Ethno- and Cultural Psychiatry. In H. Callan (Ed.), *The International Encyclopedia of Anthropology* (pp. 1–11). John Wiley & Sons, Ltd.

Kirmayer, L. J., & Ban, L. (2013). Cultural Psychiatry: Research Strategies and Future Directions. In *Cultural Psychiatry* (pp. 97–114). https://doi.org/10.1159/000348742

Kirmayer, L. J., & Bhugra, D. (2009). Culture and Mental Illness: Social Context and Explanatory Models. In I. M. Salloum & J. E. Mezzich (Eds.), *Psychiatric Diagnosis* (pp. 29–40). John Wiley & Sons, Ltd.

Kirmayer, L. J., & Crafa, D. (2014). What kind of science for psychiatry? *Frontiers in Human Neuroscience*, *8*, 435.

Kirmayer, L. J., Gomez-Carrillo, A., & Veissière, S. (2017). Culture and depression in global mental health: An ecosocial approach to the phenomenology of psychiatric disorders. *Social Science & Medicine*, *183*, 163–168.

Kirmayer, L. J., & Jarvis, G. E. (2007). Depression across cultures. In Stein, D. J., Kupfer, D. J., & Schatzberg, A. F. (Eds.). *Textbook of Mood Disorders* (pp. 699-715). American Psychiatric Publishing.

Kirmayer, L. J., & Ramstead, M. J. D. (2017). Embodiment and enactment in cultural sychiatry. In Durt. C., Fuchs, T. & Tewes. T. (Ed.), *Embodiment, Enaction, and Culture* (pp. 397–422). MIT Press.

Kirmayer, L. J., & Ryder, A. G. (2016). Culture and psychopathology. *Current Opinion in Psychology*, *8*(Supplement C), 143–148.

Kirmayer, L. J., & Sartorius, N. (2007). Cultural models and somatic syndromes. *Psychosomatic Medicine*, *69*(9), 832–840.

Kirmayer, L.J., Worthman, C., Kitayama, S. (2020a). Introduction: Co-constructing culture, mind and brain In Kirmayer, L.J., Worthman, C., Kitayama, S., Lemelson, R. & Cumming (Ed.), *Culture, Mind and Brain: Emerging Concepts, Models, Applications* (pp. 1-49). Cambridge University Press.

Kirmayer, L.J., Worthman, C., Kitayama, S. (2020b). Epilogue: Interdisciplinarity in the study of culture, mind and brain. In Kirmayer, L.J., Worthman, C., Kitayama, S., Lemelson, R. & Cumming (Ed.), *Culture, Mind and Brain: Emerging Concepts, Models, Applications* (pp. 494–512). Cambridge University Press.

Kirmayer, L. J., Worthman, C. M., Kitayama, S., Lemelson, R., & Cummings, C. (2020). *Culture, Mind, and Brain: Emerging Concepts, Models, and Applications*. Cambridge University Press.

Kirmayer, L. J., & Young, A. (1998). Culture and somatization: clinical, epidemiological, and ethnographic perspectives. *Psychosomatic Medicine*, *60*(4), 420–430.

Kirmayer, L. J., & Young, A. (1999). Culture and context in the evolutionary concept of mental disorder. *Journal of Abnormal Psychology*, *108*(3), 446–452.

Kluger, M. J. (1979). *Fever, Its Bi- ology, Evolution, and Function*. Princeton University Press.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Laland, K. N., Uller, T., Feldman, M. W., Sterelny, K., Müller, G. B., Moczek, A., Jablonka, E., & Odling-Smee, J. (2015). The extended evolutionary synthesis: its structure, assumptions and predictions. *Proceedings. Biological Sciences / The Royal Society*, *282*(1813), 20151019.

Latour, B. (2000). *Pandora's hope : essays on the reality of science studies*. Harvard University Press.

Lawson, R. P., Rees, G., & Friston, K. J. (2014). An aberrant precision account of autism. *Frontiers in Human Neuroscience*, *8*, 302.

Limanowski, J., & Friston, K. (2018). "Seeing the Dark": Grounding Phenomenal Transparency and Opacity in Precision Estimation for Active Inference. *Frontiers in Psychology*, *9*, 643.

Lyubomirsky, S., & Nolen-Hoeksema, S. (1993). Self-perpetuating properties of dysphoric rumination. *Journal of Personality and Social Psychology*, *65*(2), 339–349.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about Mechanisms. *Philosophy of Science*, *67*(1), 1–25.

Mayes, R., & Horwitz, A. V. (2005). DSM-III and the revolution in the classification of mental illness. In *Journal of the History of the Behavioural Sciences* (Vol. 41, Issue 3, pp. 249–267). https://doi.org/10.1002/jhbs.20103

McAllister, A., Fritzell, S., Almroth, M., Harber-Aschan, L., Larsson, S., & Burström, B. (2018). How do macro-level structural determinants affect inequalities in mental health? - a systematic review of the literature. *International Journal for Equity in Health*, *17*(1), 180.

McDonnell, C. G., Valentino, K., & Diehl, J. J. (2017). A developmental psychopathology perspective on autobiographical memory in autism spectrum disorder. *Developmental Review: DR*, *44*, 59–81.

Mealey, L. (1995). The sociobiology of sociopathy: An integrated evolutionary model. *The Behavioural and Brain Sciences*, *18*(3), 523–541.

Mercier, H., & Sperber, D. (2017). *The Enigma of Reason*. Harvard University Press.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, *16*(1), 72–80.

Murphy, D. (2005). Can evolution explain insanity? *Biology & Philosophy*, *20*(4), 745–766.

Murphy, D., & Woolfolk, R. L. (2000). The harmful dysfunction analysis of mental disorder. *Philosophy, Psychiatry, & Psychology: PPP*, *7*(4), 241–252.

Nelson, K., & Fivush, R. (2004). The emergence of autobiographical memory: a social cultural developmental theory. *Psychological Review*, *111*(2), 486–511.

Nesse, R. M. (1999). What Darwinian medicine offers psychiatry. *Evolutionary Medicine*.

Nesse, R. M. (2001). The smoke detector principle. Natural selection and the regulation of defensive responses. *Annals of the New York Academy of Sciences*, *935*, 75–85.

Nesse, R. M. (2002). Evolutionary biology: a basic science for psychiatry. *World Psychiatry: Official Journal of the World Psychiatric Association* , *1*(1), 7–9.

Nesse, R. M. (2007). Evolution is the scientific foundation for diagnosis: psychiatry should use it. *World Psychiatry: Official Journal of the World Psychiatric Association* , *6*(3), 160–161.

Nesse, R. M. (2011). Ten questions for evolutionary studies of disease vulnerability. *Evolutionary Applications*, *4*(2), 264–277.

Nesse, R. M. (2017). Evolutionary foundations for psychiatric research and practice. In B. J. Sadock, V. A. Sadock, & P. Ruiz (Ed.), *Kaplan & Sadock's comprehensive textbook of psychiatry* (Vol. 10, pp. 769–780).

Nettle, D. (2004). Evolutionary origins of depression: a review and reformulation. *Journal of Affective Disorders*, *81*(2), 91–102.

Nettle, D. (2006). The evolution of personality variation in humans and other animals. *The American Psychologist*, *61*(6), 622–631.

Ng, A., & Jordan, M. (2002). On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes. In T. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems* (Vol. 14). MIT Press.

Nicholson, D. J. (2012). The concept of mechanism in biology. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 152–163.

Nolen-Hoeksema, S. (1991). Responses to depression and their effects on the duration of depressive episodes. *Journal of Abnormal Psychology*, *100*(4), 569–582.

Paris, J., & Kirmayer, L. J. (2016). The National Institute of Mental Health Research Domain Criteria. In *Journal of Nervous & Mental Disease* (Vol. 204, Issue 1, pp. 26–32). https://doi.org/10.1097/nmd.0000000000000435

Parr, T., Benrimoh, D. A., Vincent, P., & Friston, K. J. (2018). Precision and False Perceptual Inference. *Frontiers in Integrative Neuroscience*, *12*, 39.

Parr, T., & Friston, K. J. (2019). Attention or salience? *Current Opinion in Psychology*, *29*, 1–5.

Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active Inference: The free-energy Principle in Mind, Brain, and Behaviour*. MIT Press.

Parr, T., Rees, G., & Friston, K. J. (2018). Computational Neuropsychology and Bayesian Inference. *Frontiers in Human Neuroscience*, *12*, 61.

Paton, B., Hohwy, J., & Enticott, P. G. (2012). The rubber hand illusion reveals proprioceptive and sensorimotor differences in autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *42*(9), 1870–1883.

Pickering, A. (1995). *The mangle of practice: time, agency, and science*. University of Chicago Press.

Queller, D. C. (2014). Joint phenotypes, evolutionary conflict and the fundamental theorem of natural selection. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *369*(1642), 20130423.

Ramstead, M. J., Constant, A., Badcock, P. B., & Friston, K. J. (2019). Variational ecology and the physics of sentient systems. *Physics of Life Reviews*. https://doi.org/10.1016/j.plrev.2018.12.002

Ramstead, M. J. D., Badcock, P. B., & Friston, K. J. (2018). Answering Schrödinger's question: A free-energy formulation. *Physics of Life Reviews*, *24*, 1–16.

Ramstead, M. J. D., Kirchhoff, M. D., Constant, A., & Friston, K. J. (2019). Multiscale integration: beyond internalism and externalism. *Synthese*. https://doi.org/10.1007/s11229-019-02115-x

Ramstead, M. J. D., Veissière, S. P. L., & Kirmayer, L. J. (2016). Cultural affordances: scaffolding local worlds through shared intentionality and regimes of attention. *Frontiers in Psychology*, *7*, 1090.

Raubenheimer, D., Machovsky-Capuska, G. E., Gosby, A. K., & Simpson, S. (2015). Nutritional ecology of obesity: from humans to companion animals. *The British Journal of Nutrition*, *113 Suppl*, S26–S39.

Richerson, P. J. (2018). An integrated bayesian theory of phenotypic flexibility. *Behavioural Processes*. https://doi.org/10.1016/j.beproc.2018.02.002

Rietveld, E., & Kiverstein, J. (2014). A rich landscape of affordances. *Ecological Psychology: A Publication of the International Society for Ecological Psychology*, *26*(4), 325–352.

Riggs, J. E. (1993). Stone-age genes and modern lifestyle: evolutionary mismatch or differential survival bias. *Journal of Clinical Epidemiology*, *46*(11), 1289–1291.

Roepstorff, A., Niewöhner, J., & Beck, S. (2010). Enculturing brains through patterned practices. *Neural Networks: The Official Journal of the International Neural Network Society*, *23*(8-9), 1051–1059.

Rose, H., & Rose, S. (2010). *Alas Poor Darwin: Arguments Against Evolutionary Psychology*. Random House.

Rosenberg, A. (2018). Making mechanism interesting. *Synthese*, *195*(1), 11–33.

Rosenberg, A. (2020). *Reduction and Mechanism*. Cambridge University Press.

Rude, S. S., Valdez, C. R., Odom, S., & Ebrahimi, A. (2003). Negative Cognitive Biases Predict Subsequent Depression. *Cognitive Therapy and Research*, *27*(4), 415–429.

Schwartz, P. H. (2007). Defining Dysfunction: Natural Selection, Design, and Drawing a Line. *Philosophy of Science*, *74*(3), 364–385.

Sigmund, K., & Nowak, M. A. (1999). Evolutionary game theory. *Current Biology: CB*, *9*(14), R503–R505.

Sloan, D. M., Strauss, M. E., Quirk, S. W., & Sajatovic, M. (1997). Subjective and expressive emotional responses in depression. *Journal of Affective Disorders*, *46*(2), 135–141.

Smoller, J. W., & Stein, M. B. (2018). Precision Psychiatry-Yes, But Precisely What?-Reply. *JAMA Psychiatry* , *75*(12), 1303.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, *46*(4), 1004–1017.

Stevens, A., & Price, J. (2015). *Evolutionary psychiatry: A new beginning*. Routledge.

Teasdale, J. D. (1983). Negative thinking in depression: Cause, effect, or reciprocal relationship? *Advances in Behaviour Research and Therapy*, *5*(1), 3–25.

Tononi, G., & Cirelli, C. (2006). Sleep function and synaptic homeostasis. *Sleep Medicine Reviews*, *10*(1), 49–62.

Troisi, A. (2005). The concept of alternative strategies and its relevance to psychiatry and clinical psychology. *Neuroscience and Biobehavioral Reviews*, *29*(1), 159–168.

Troisi, A. (2006). Adaptationism and medicalization: The Scylla and Charybdis of Darwinian psychiatry. *The Behavioural and Brain Sciences*, *29*(4), 422–423.

Troisi, A. (2020). Social stress and psychiatric disorders: Evolutionary reflections on debated questions. *Neuroscience and Biobehavioral Reviews*, *116*, 461–469.

Tseng, W.-S. (2001). *Handbook of Cultural Psychiatry*. Academic Press.

Tsou, J. Y. (2021). *Philosophy of psychiatry*. Cambridge University Press.

Van de Cruys, S., Evers, K., Van der Hallen, R., Van Eylen, L., Boets, B., de-Wit, L., & Wagemans, J. (2014). Precise minds in uncertain worlds: predictive coding in autism. *Psychological Review*, *121*(4), 649–675.

Varga, S. (2012). Evolutionary psychiatry and depression: testing two hypotheses. *Medicine, Health Care, and Philosophy*, *15*(1), 41–52.

Vasil, J., Badcock, P. B., Constant, A., Friston, K., & Ramstead, M. J. D. (2020). A World Unto Itself: Human Communication as Active Inference. *Frontiers in Psychology*, *11*, 417.

Veissière, S. P. L. (2016). Varieties of Tulpa experiences: the hypnotic nature of human sociality, personhood, and interphenomenality. In: A. Raz & M Lifhsitz (Eds.) *Hypnosis and Meditation: Towards an Integrative Science of Conscious Planes* (pp. 55–76). Oxford University Press.

Veissière, S. P. L., Constant, A., Ramstead, M. J. D., Friston, K. J., & Kirmayer, L. J. (2020). Thinking through other minds: A variational approach to cognition and culture. *The Behavioural and Brain Sciences*, *43*. https://doi.org/10.1017/S0140525X19001213

Veissière, S. P. L., & Stendel, M. (2018). Hypernatural Monitoring: A Social Rehearsal Account of Smartphone Addiction. *Frontiers in Psychology*, *9*, 141.

Wakefield, J. C. (1992a). The concept of mental disorder. On the boundary between biological facts and social values. *The American Psychologist*, *47*(3), 373–388.

Wakefield, J. C. (1992b). Disorder as harmful dysfunction: a conceptual critique of DSM-III-R's definition of mental disorder. *Psychological Review*, *99*(2), 232–247.

Wakefield, J. C. (2005). Biological Function and Dysfunction. In D. M. Buss (Ed.), *The handbook of evolutionary psychology , (pp* (Vol. 1028, pp. 878–902). John Wiley & Sons, Inc., xxv.

Young, A., & Breslau, N. (2016). What Is "PTSD"? The Heterogeneity Thesis. In D. E. Hinton & B. J. Good (Eds.), *Culture and PTSD*. University of Pennsylvania Press.

Zhang, X.-S., & Hill, W. G. (2005). Genetic variability under mutation selection balance. *Trends in Ecology & Evolution*, *20*(9), 468–470.

Conclusion to chapter 3

One worry that I had when I started this project was that there wouldn't be one coherent way to discuss evolutionary, cultural, and computational psychiatry all at once. I feared that combining these approaches — by foregrounding what they share — would lead to a combinatorial explosion of problems; each approach having its own series of problems, which, when piled up, would likely create new problems. Hence, I had, under the advice of my mentors, notably of Paul Griffiths, to let go of some discussions that were central to the approaches taken individually. One of those problems is the problem of evolutionary function in evolutionary psychiatry. In chapter 3, this problem is briefly introduced as one of the two problems faced by the concept of mental disorders as harmful dysfunctions in evolutionary psychiatry. The problem is that of finding an evolutionary standard against which to evaluate dysfunctionality. There are typically two types of accounts to solve the problem of evolutionary function. The first is the *biostatistical* account, and the second is the *selected effect* account (Griffiths & Matthewson, 2018). These are sometimes called the "non-aetiological" and "aetiological" accounts (Schwartz, 2007). To conclude this chapter, I would like to revisit that problem in more details.

The biostatistical account (Boorse, 1977) articulates three criteria to trace the boundary between the normal and the pathological: (i) physiological function, or the ability to contribute to fitness; (ii) the reference class within the species (e.g., defined by age group and sex); and (iii) statistical normality, or frequency in the distribution of the trait in the reference class (Giroux, 2015). The biostatistical account is not historical, as the physiological function refers to the ability of the trait to contribute to current and future survival and reproductive success. The biostatistical account does not need to rely on the idea that the trait has been designed for a certain purpose (which it would fail in the case of a pathology), nor, consequently, does it need to appeal to a comparison with the population-level fitness. To evaluate its first criterion (i.e., does the trait fulfil its physiological goal?), all the biostatistical account uses is a comparative analysis of net fitness (i.e., realized) between the putatively pathological trait and its most frequent configuration or expression in the reference class (criteria 2 and 3). For a critical analysis see (Matthewson & Griffiths, 2017).

Conversely, the selected effect account of function (Godfrey-Smith, 1994; Griffiths, 1993; Millikan, 1984) evaluates the normality of a trait in terms of proper functioning. A proper function of a trait is that which produces an effect for which the trait gained a competitive advantage in terms of reproductive success in the past (i.e., was selected for), thereby explaining its current frequency in the population. The selected account is historical in the sense that to make a claim about the normal and the pathological, one must consider the manner in which a trait performed its function in the past, and in so doing, explain the current frequency of that trait. If the trait is such that it would perform its function in a suboptimal fashion (in the environment in which it has been designed to operate in) compared to the (weighted average) fitness at the level of the population, then, the trait is normatively described as dysfunctional or pathological. Assuming that a trait has been designed by natural selection to perform a certain function, failing to perform that function is sufficient to claim that that trait is pathological.

The model we present in chapter 3 sits well with the selected effect account, and thus should underwrite the way one selects adaptive priors to parameterize an ECC model.

Under the selected effect account, pathologies must be studied in light of the multiple factors involved in the developmental pathway that leads to the pathological state of affairs—a pathway whose function is characterized by the assumption of evolutionary design. Accordingly, under that view, functional failure may be due to a broken biological mechanism, but also to environmental conditions that are 'abnormal' relative to the context within which the trait was designed to operate (that is, selected for). Alternatively, functional failure may simply be due to finding oneself in an inhospitable environment – which might not be abnormal (i.e., unusual) as per the biostatistical account, but nonetheless deleterious to functional performance. Functional failure may also ensue because of the lack of developmental experiences required for the normal developmental trajectory expected under the sort of life cycle the agent has been designed to follow (Matthewson & Griffiths, 2017). The selected account thus admits many ways in which a function can fail, the diagnosing of which requires an evaluation of the current and past environment. It equally admits the possibility that a trait that may initially look statistically pathological might in fact function normally (i.e., as per its designed), though under abnormal environmental conditions (e.g., as in the case of mismatch).

Under the selected effect view, the 'naïve' evolutionary rationale (Kirmayer & Young, 1999) is simply employed as a backstory to make sense of why the trait we observe now strikes us as pathological, given the person's affliction. The evolutionary rational functions as an *epistemic standard* (e.g., a position from which to start thinking) that forces us to think backward through the evolutionary history of the trait to make sense of how the trait operates now. Accordingly, the selected effect account does not need to refer to evolution (i.e., survival and fitness) as a norm to account for the dysfunctional nature of a trait and requires an assessment of the context (e.g., cultural context) to evaluate dysfunctions. Conversely, the evolutionary rationale under the biostatistical view takes reproductive success and survival as *normative standards* (e.g., criteria for cut-offs to distinguish pathology) that license assessments of dysfunction limited to current observations.

Chapter 4 will present a model of symptoms of normative depression and intervention that can be viewed as implementing the conceptual model proposed in chapter 3. Crucially, it should be noted that the evolutionary rationale that we follow to parameterize the model — that of the Evolutionary System Theory of Depression — aligns with the selected effect account, not the biostatistical account. This rationale does not describe normatively what makes depression a disorder from an evolutionary point of view, but rather allows one to make sense of why certain responses to current environmental contexts may lead to affliction characteristic of major depressive depression, despite these responses being possibly adaptive under different contexts. While this may sound like a rather abstract theoretical preference, opting for a selected effect account over a biostatistical perspective would have very concrete implication in terms of fitting the evolutionary prior parameter to a clinical population in the context of an empirical study seeking to reproduce our simulated results.

Under a biostatistical view, assuming that the parameterization of the evolutionary prior is what precipitates the individual into a state of normative depression, the sampled clinical population for the study would have to be selected based on criteria that would reflect pathological simulated parameters: in the example we will

present, these have to do with degrees of preference for social partners. More generally, this requires defining what a pathological parameter is in the first place, which brings us back to the line drawing problem. Conversely, under a selected effect account, which allows for non-normative evolutionary thinking about mental disorders, one would not need to define a depression-specific parameterization of the evolutionary prior. Rather, one could provide a generic model whose development leads to symptoms of depression due to a normal evolutionary prior that operates under abnormal environmental conditions that ought to be modelled as well, which is what we do in Chapter 4. Empirically testing such a model would not require one to look for selection criteria reflecting some arbitrarily defined pathological parameterization of the evolutionary prior. Rather, it would require selecting the population based on criteria that reflect the way empirical (i.e., entirely learnable) priors can yield pathological outcomes, e.g., based on the psychosocial background of the person.

Chapter 4: Why Depressed Mood is Adaptive: A Numerical Proof of Principle for an Evolutionary Systems Theory of Depression

Introduction to chapter 4

Chapter 4 is entitled *Why Depressed Mood is Adaptive: A Numerical Proof of Principle for an Evolutionary Systems Theory of Depression*. Chapter 4 provides a simulation study of anhedonia and social withdrawal, which are features of depressed mood associated with normative depression. Explicitly, the proposed simulation is meant to work as a proof of principle for the evolutionary systems theory (EST) of depression proposed by my colleague and mentor, Paul Badcock. The EST of depression suggests that normative depressive symptoms result from the spiralling of normal responses to an increase in social network uncertainty (e.g., decreased reliability of partners' expected availability); responses which, under normal environmental circumstances, should increase interpersonal support via social signalling. The simulation induces severe depression in the simulated agent and returns the agent to normal mood via synthetic social and pharmacological interventions.

Implicitly, the simulation of chapter 4 functions as a "proof of principle" for the implementation of the model developed in chapter 3; a proof that concrete research outputs can be generated based on the ECC conceptual model of chapter 3. As mentioned in the introduction, the model of chapter 3 is a conceptual model that can serve as a means of organizing knowledge so as to guide scientific practice. In chapter 4, the proposed model of the EST of depression results from organizing knowledge and modes of reasoning in evolutionary, cultural, and computational approaches to psychiatry in order to observe, *in silico*, theoretical claims of the EST. Following chapter 3, the model of chapter 4 simulates the causes of pathological depression as developmentally aggravated vulnerabilities viewed as a proximate cause and as behavioural patterns causing psychological distress through looping interactions with the environment and as atypically learned model parameters. The return to normal mood, then, is achieved by returning the model parameters to values that would yield healthy behaviour.

1 Introduction

It has recently been proposed that depressed mood reflects an adaptive, socially risk-averse psychobiological strategy that preserves social relationships (i.e., inclusion) when there is evidence for maladaptive instability in interpersonal exchanges (Badcock et al., 2017). This perspective follows an evolutionary systems theory (EST) of human biobehaviour called the hierarchically mechanistic mind, which combines insights drawn from research in psychology with the computational resources borrowed from the theory of active inference in theoretical neurobiology (Badcock, Friston, & Ramstead, 2019). This model rests on two fundamental claims. The first conforms to the theory of active inference in theoretical neurobiology by suggesting that the brain comprises hierarchically organized neurocognitive mechanisms that reduce the dispersion or decay of our sensory and phenotypic states—by generating action-perception cycles that minimize surprising exchanges with the world. The second claim ensues from an embodied perspective on neural form and function—that accommodates the broader evolutionary, developmental, and real-time processes that act on human phenotypes. The implication here is that to understand a phenotypic trait, we need approaches that synthesise findings from diverse fields of inquiry to explain both why that trait is adaptive, along with how it emerges from the nested dynamics across different timescales. In this spirit, the current study provides proof of principle for the EST of depression, using simulations of active inference. We conclude by discussing the clinical implications of our model.

Our proof of principle integrates two major schools of thought. The first is rooted in evolutionary psychological approaches to depression, rallied around the social risk hypothesis (SRH) proposed by Allen and Badcock (Allen & Badcock, 2003). Psychological symptoms of depression include feelings of sadness, emptiness, and hopelessness, along with systematic disinterest in activities (i.e., anhedonia), feelings of worthlessness, and inappropriate guilt. Typically, a diagnosis of depression is made when symptoms have been present for at least 14 days (American Psychiatric Association, 2013). Two important symptoms of depression are anhedonia and social withdrawal: the latter is commonly observed in depression as a clinical correlate of anhedonia, but is not a formal criterion (Buckner et al., 2008). Evolutionary models of depression explain the maintenance of genetic vulnerabilities to depressive symptomatology in terms of the selective advantage of these vulnerabilities in ancestral environment (R. M. Nesse, 1990)[i].

The adaptive properties of depression are thought to be restricted to the relatively transient, normative depressed mood states that we all experience from time to time, while more severe manifestations, like those observed in major depressive disorder, reflect a dysregulation of our species-typical capacity for mood variation (Nettle, 2004). The SRH suggests that depressive symptoms might have been selected as a strategy that prevents the deterioration of interpersonal relationships. Low mood reduces one's propensity for social risk-taking, and increases implicit signalling for social support, which reduces competitive encounters (see box 1 for background). Clinical depression occurs when this sequence becomes maladaptive; specifically, when it does not lead to a resumption of normal mood. This may be due to neurobiological or psychological deficits that maintain increased sensitivity to social instability, or to instabilities or the absence of support in the proximal environment towards which the depressed individual reacts. The increased sensitivity to social

instability constitutes the basis of the neurocognitive processes leading to depressed mood, upon which selection can act (for a review see (Badcock, Friston, Ramstead, et al., 2019)).

Box 1 Evolution and depression

There are three general classes of evolutionary models of normative depressed mood. The first of these – resource conservation views – claims that depressive symptoms, such as learned helplessness, are a response to a low positive reward rate and insufficient control over reward and punishment (Randolph M. Nesse, 2000)[ii], and unobtainable incentives or goals (Klinger, 1975). Low appetitive functions (e.g., anhedonia) allow the individual to fine-tune resource allocation by precluding investment in poor pay-off activities. The second class refers to the social competition model, which claims that social status (e.g., rank and position in the social group) positively correlates with access to resources that enhance reproductive success. Depressive symptoms such as social withdrawal remove the individual from conflicts and other status-impairing situations that would negatively impact their social rank (Gilbert, 1997; Price, 1967). The third, attachment model claims that given the delayed maturation of human infants (Hrdy, 2011), offspring survival necessitates intensive parental and alloparental investment. Behaviours designed to maintain proximity to caregivers are instigated when significant affectional bonds are threatened. In the face of precarious interpersonal relationships, depressive symptoms should promote help-seeking and inhibit exploratory behaviour and risk-taking, thereby maintaining relationships with the proximal familial environment, while avoiding the deterioration of current social bonds (Ingram et al., 1998). Standing alone, it has been argued that these models cannot account for the full scope of the depressive phenotype (Allen & Badcock, 2006). Darwinian models do not provide an explanation of the underlying mechanisms upon which selection can act. For instance, how is 'resource reallocation', 'preventive withdrawal from the social environment', or 'familial bond strengthening' implemented mechanistically? This is a problem, since the unit of selection is never a complex behavioural trait, but rather some (epi)genetic dispositions to express such traits. The Social Risk Hypothesis (SRH) was proposed to provide a solution to these issues and has since been developed into a neurobiologically plausible and empirically tractable mechanistic explanation for depressive phenomena (Badcock et al., 2017). According to this view, normative levels of depressed mood reduce the probability of deleterious social outcomes via three broad classes of action: (1) depression increases individuals' cognitive sensitivity to environmental cues of social risk or instability; (2) it reduces their behavioural propensity for social risk-taking; and (3) it generates signalling behaviours (e.g., reassurance seeking, crying, gaze aversion) that attract social support and defuse aggressive or competitive encounters.

The second root of our simulation centres upon formal, testable models of depressive phenomena that borrow from the principles of computational psychiatry (Huys, Guitart-Masip, et al., 2015). Accordingly, a key aspect of the work reported in this paper is the attempt to model social inference – as it relates to the phenomenology of depression – from first principles. This is challenging, because of the many aetiological factors that underwrite the psychopathology and pathophysiology of depression. We try to formalize the normative aspects of depression as a Bayes optimal response to inference in the prosocial world, while considering both social and pharmacological interventions. To our knowledge, this is the first modelling work that addresses the interaction between social factors and pharmacotherapy within the same formalism. In this sense, the simulations reported here also provide a proof of principle for a model of the effects of drug treatment on neuronal computations that underlie belief updating and behaviour in depression. In brief, we make three basic assumptions that allow us to characterize the effect of drug treatment on social inference and subsequent behaviour. First, both inference and learning conform to the same (ideal Bayesian observer) principles of active inference; namely, belief updating and experience-dependent plasticity both optimize a

variational free-energy bound on (log) model evidence or marginal likelihood (Friston et al., 2016). Second, pharmacotherapy motivates neuromodulatory effects that, computationally, change the precision of sub-personal probabilistic beliefs (i.e., prior beliefs about states of affairs in the world or likelihood mappings between causes and consequences) (Parr et al., 2018). Finally, one cannot ignore the reciprocal coupling between an agent and her (prosocial) environment when modelling interpersonal exchanges. This requires an explicit consideration of how environmental (prosocial) contingencies respond to an agent's behaviour (Badcock et al., 2017) (see box 2 for background). Here, the embedded aspect of interventions on the social environment (Bruineberg et al., 2018; Constant et al., 2018) was modelled by an increase in social reliability following patterns of behaviour that can be construed as social signalling. Our hope was to show that functional responses to social adversity use the same inferential mechanisms seen in pathological depression – and that psychopathology can be remediated by a combination of social support and drug therapy.

Box 2 Computation, inference and depression

Computational phenotyping is a method in computational psychiatry to test hypotheses about the neurophysiology of mental disorders to inform nosology and suggest treatment approaches (Corlett & Fletcher, 2014). A computational phenotype refers to the set of measurable features of an agent; often described in terms of the 'priors' and 'likelihood mappings' of a generative model used by subjects for perception and decision-making (Schwartenbeck & Friston, 2016). The associated parameters of generative models show variation across the population upon which selection acts (Montague et al., 2012). This formal approach to phenotyping effectively reduces the phenotype to some formal priors – or prior beliefs – that personalize a (generative) model that people use to predict and interact with their (physical, physiological, prosocial or cultural) econiche. These interactions are usually cast in terms of (Bayesian) belief updating under a generative model that characterizes a given phenotype. Computationally, prior probability distributions can take numerous forms (e.g., normal, Dirichlet, delta). This form depends on the state space being modelled (e.g., discrete, as in this paper, or continuous). Priors can also have different 'temporal' scales, relative to the scale at which (Bayesian) belief updating unfolds. 'Adaptive priors' are sculpted by evolutionary processes and become encoded over the course of ontogeny in the physiology and functional architectures of the brain; that is, they emerge from interactions between priors that are 'empirical' and 'evolutionary'. Empirical priors are learned over development (e.g., learned distribution of food patches), while evolutionary priors function as initial conditions that shape the learning of empirical priors (e.g., a prior preference for energy-rich food) (Friston, 2010). Technically, empirical priors arise whenever there is a hierarchical generative model. Empirical priors are the constraints offered to a lower level, from a higher level. When hierarchical models are inverted, empirical priors become informed by (empirical) data. Simulating belief updating – under a generative model – allows researchers to produce synthetic, *in silico* measurements (e.g., psychophysical and physiological responses) of the sort that are usually studied in real-world empirical contexts. The generative model can be manipulated by inducing artificial lesions in likelihood mappings, or by simulating pharmacological treatment that generally changes the priors (Parr et al., 2018). In so doing, one can generate artificial data in the context of a task that can be used in empirical studies. One can then test hypotheses by comparing artificial responses with participants' empirical data (Cullen et al., 2018). The symptoms of depression are thought to relate to deficits impacting long-term reward evaluation through the acquisition of 'pessimistic' priors that entail negatively biased learning of environmental states (Huys, Daw, et al., 2015). Priors should reflect unambiguous beliefs about the world, as well as beliefs about the relation between environmental states and observed outcomes. For instance, we cannot directly infer the mood of another solely from the sensory impressions of that person's facial expression. Rather, we must consider some prior assumptions about the person's behaviour and outcome probability over time (e.g., she usually smiles a lot, but she is not smiling today, so it is likely that something

is awry). These are empirical priors. In depression, empirical priors biasing decision-making tip the balance towards pessimistic inference, thereby leading to systematically pessimist thoughts. For instance, depressive patients form negative sentences more quickly and frequently than non-depressed controls, when presented with both optimistic and pessimistic options (e.g., in the scramble sentence test) (Hindash & Amir, 2012; Rude et al., 2003). The simulations offered in this paper try to capture this belief-based phenomenology by using synthetic agents and active inference – a generic framework for (active) Bayesian inference and planning.

Our numerical proof of principle is based on active inference for discrete states, using (Markovian) generative models (Friston, Parr, et al., 2017) – see method section. We present a series of simulations based on an augmented version of a two-armed bandit game from economics, in which the agent has to choose between a risky or safe social engagement (Schwartenbeck et al., 2019). The augmentation involved offering the agent with a cue option that indicates whether the risky arm is low-risk or high-risk (i.e., it indicates the social context); this contextual state alternates every other trial. Narratively, the cue corresponds to social media that provides information or evidence that reduces uncertainty about the prevailing social context. This means that healthy agents will systematically sample the cue to make an informed decision. As they sample their social environment, agents will learn the probability of reward afforded by choosing one of the two arms. Narratively, this relates to checking people's availability before choosing among social options.

Based on the learning process characteristic of active inference, agents exposed to social adversity (e.g., rejection by social partners) will learn the likelihood of being rejected. We simulate different phenotypes, with and without social support and pharmacotherapy, which reshape the agent's (pessimistic) prior beliefs. Our numerical analyses speak to how pharmacotherapy and social support – triggered by social signalling on social media – allows the agent to regain a normal mood. Depending on the type of intervention (social, pharmaceutical, or the lack thereof), the agent typically experiences a phase of low mood, and either spirals into persistent depression (anhedonia and social withdrawal), or returns to various levels of normal functioning. We will quantify the responses of our synthetic agent in terms of task performance and associated synthetic mood (i.e., expected reward under a given action policy) and behaviour (action selection).

Our two-arm bandit social decision-making task (see Figure 1) involves choosing among three social engagement options, which vary in their risk. The first is a 'safe' option, but with low social preference (going to a well-known friend, Rudolph, who you know can be engaged with 100% success, but won't provide the most fulfilling interaction). The second, 'risky' option has a high preference (going to see a new popular student, Caroline, whom you do not know, but were told is a lot of fun), but is risky because Caroline often forfeits, and the agent is averse to failed social encounters. On a 'good day' the agent has a 75% chance of successfully engaging Caroline, but on a 'busy day', only a 25% chance of catching her. The third, 'socially epistemic' option yields a null cost: under this option, the agent can turn to social media, to see if Caroline is having a 'busy day'. Each trial, or day, involves three time steps, the second of which is the one where the agent can check the epistemic cue.

To characterize prosocial and emotional inference that might underwrite depression, we considered belief updating and subsequent behaviour under 8 different conditions – in a three-way factorial design involving the following factors: social adversity, social support, and pharmacotherapy (see table 1).

Table 1 Interventions

Baseline	The agent performs the social decision-making task in the absence of any adversity over a period of 64 days.
Severe depression	We induce social adversity on the 28th day by changing the uncertainty of social outcomes. The agent is now rejected by Rudolph (always) and by Caroline on a 'bad day'. On a good day, the odds are inverted, such that Caroline is likely to afford a negative outcome. In other words, there is a flip in contingencies of the social environment.
Social support	We introduce social support on the 30th day, which reduces uncertainty about the outcomes of social encounters – and therefore resolves social adversity. This is modelled as an increase in Rudolph and Caroline's reliability, which is increased when the agent forages for information on social media. Narratively, this could be interpreted as the agent signalling (implicitly or explicitly) to Caroline and Rudolph that they should be more consistent. Recovery thus depends on the sensitivity of the social environment and on how often the agent consults social media.
Pharmacotherapy	First-line pharmacotherapy typically employs either selective serotonin or norepinephrine reuptake inhibitors, and sometimes mixed serotonin or norepinephrine reuptake inhibitors (e.g., venlafaxine and duloxetine); the latter usually being used in patients who do not respond to serotonin reuptake inhibitors (Harmer et al., 2017). We simulate two types of synthetic pharmacotherapy: one motivated by serotonin and the other by norepinephrine. We assume, based on (Harmer et al., 2017), that serotonin upregulates prior expectations about initial states (i.e., increases the perceived probability of Caroline showing up), whereas noradrenaline introduces uncertainty about state transitions. Noradrenaline entails an overall loss of precise belief-updating during planning, a loss which underwrites the exploration of states that may lead to social reward. Condition 3 involves both noradrenaline and serotonin, condition 4 noradrenaline only, and condition 5 serotonin only.
Social support and pharmacotherapies	Condition 6 involves social support and both antidepressants; condition 7 involves support and noradrenaline only; and condition 8, support and serotonin only.

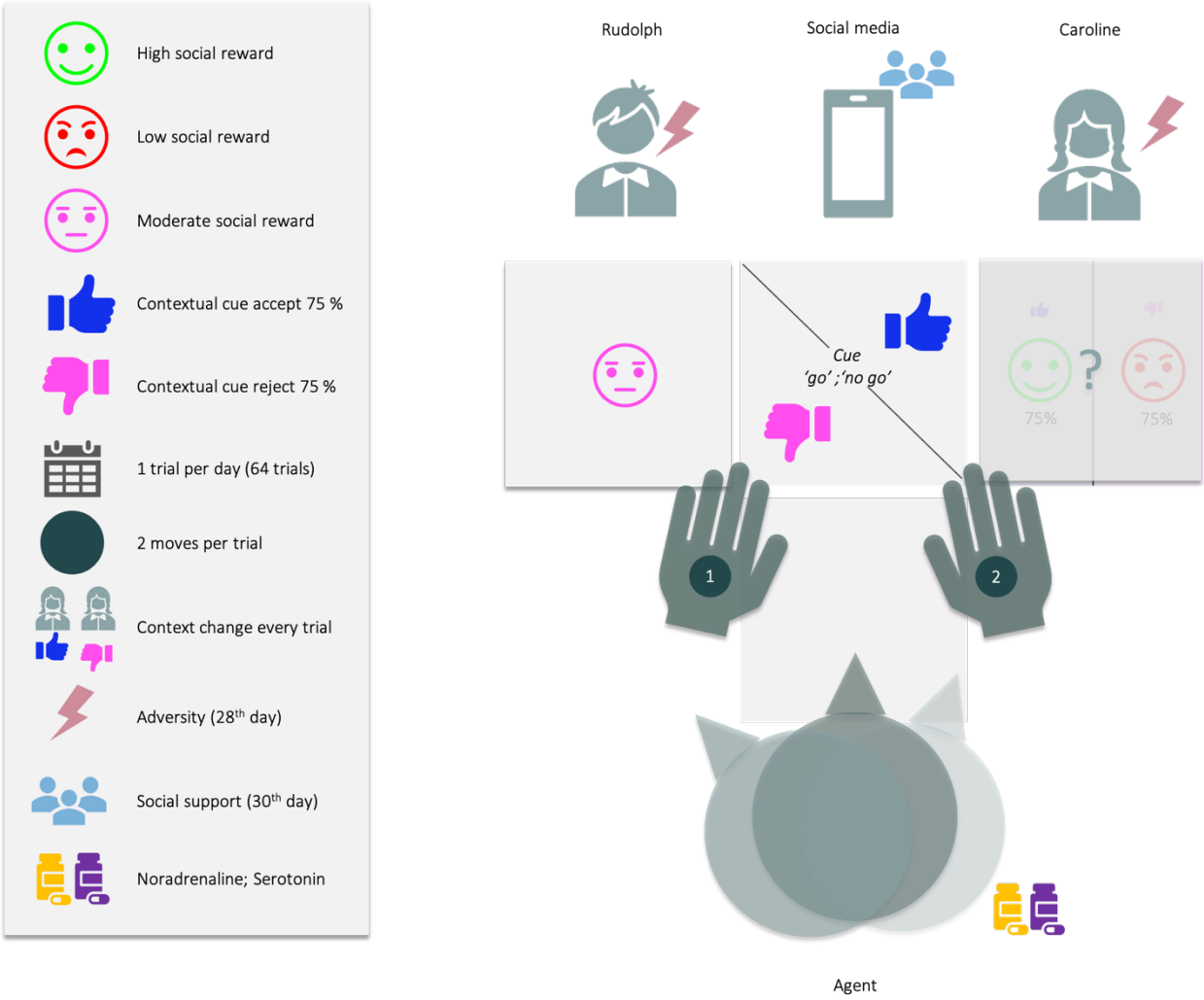


Figure 1. Narrative description of the social decision-making task. Over 64 days, the challenge is to maximize social encounters with Caroline. The agent has two moves (Caroline and Rudolph are both absorbing states, meaning that once the agent reaches them, it must stay there). For instance, on the first move, the agent can solicit information about Caroline's availability by going on social media, and then, on the second move, decide where to go.

2 Methods and materials

Active inference is a Bayesian framework that only uses local information (i.e., there is no external supervision) for belief-updating, in order to ensure biological plausibility. Markov Decision Processes (MDPs) can be used to simulate how agents infer which discrete hidden states (*s*) of the world provide the best explanation of observed sensory outcomes (*o*), under a given generative model. To generate predictions of sensory outcomes, an agent needs prior expectations about initial hidden states (an initial state prior, **D**), how states generate sensory outcomes (sensory mapping, **A**), and how states evolve over time (state transitions, **B**). The agent can infer states of the world by minimizing the discrepancy between predicted and observed outcomes (a.k.a., variational free energy), or equivalently, by maximizing Bayesian model evidence. For mathematical details, see (Parr & Friston, 2017).

When expectations of hidden states are conditioned upon the agent's plan or policy (as encoded in the policy dependent **B** matrices), one has a generative model of action (see Figure 2). Without an external referee to say what is right or wrong, the agent will need to: (i) predict her course of action, based on the succession of states, expected under each policy; and (ii) select her action based on (posterior) beliefs about the best policy. To that end, we equip the agent with (self-referential) prior beliefs that are biased towards policies with stronger expected model evidence or, equivalently, lower expected free energy, **G**.

Mathematically, expected free-energy can be decomposed into pragmatic and epistemic components for any given policy. On the one hand, pragmatic value (i.e., exploitation) biases policy selection towards obtaining preferred sensory outcomes (evolutionary prior preferences, **C**), much like utility in reinforcement learning. On the other hand, epistemic value (i.e., exploration) biases policy selection towards the (expected) minimization of uncertainty about states of the world (a.k.a., artificial curiosity).

Uncertainty can be over beliefs about current hidden states or model parameters (as quantified in free-energy **F**) or over beliefs about future hidden states and their associated outcomes under a given policy (as quantified in expected free-energy $\mathbf{G}_{\pi}$). In active inference, the latter guides action selection and can be decomposed in three distinct sources of uncertainty: (i) expected ambiguity, or anticipated uncertainty about hidden states (e.g., "how certain will I be about Caroline's mood, given that I check social media?"), (ii) expected risk, or the anticipated uncertainty about whether future outcomes will align with preferences **C** (e.g., "how certain will I be that I obtain preferred outcomes, given that I visit Caroline?"), and (iii) the anticipated uncertainty about Dirichlet parameters of the likelihood mapping **A** (e.g., "how much might I learn about state-outcome mappings if I visit Rudolf?") (Kaplan & Friston, 2018). Each of these sources of uncertainty can be manipulated directly with interventions on the model. Here, we focus on direct intervention on salience—via serotonergic and noradrenergic manipulation of initial states and state transitions—and on the indirect manipulation of extrinsic value via the manipulation of social observations, or outcomes (fig. 2). Thus, in our simulations, our agent will have a double incentive for social engagement: (i) fulfilling preferences for positive social outcomes; and (ii) the natural drive towards resolving her uncertainty over the various beliefs she has about the social world (c.f., curiosity about a new acquaintance). Crucially, it is this double incentive that we exploit to formalize the behavioural dynamics envisaged by the EST of depression; the first incentive relating to 'evolutionary' prior preferences for high social reward, and the second incentive relating to 'developmental' learning.

We limit the notion of social engagement to face-to-face encounters with Rudolph or Caroline.

The software to simulate belief updating and action selection, based on the specification of any generative model (as the one specified in Figure 2), is freely available as part of the academic software SPM; specifically, the Matlab routine spm_MDP_VB_X.m (https://www.fil.ion.ucl.ac.uk/spm/software/spm12/).

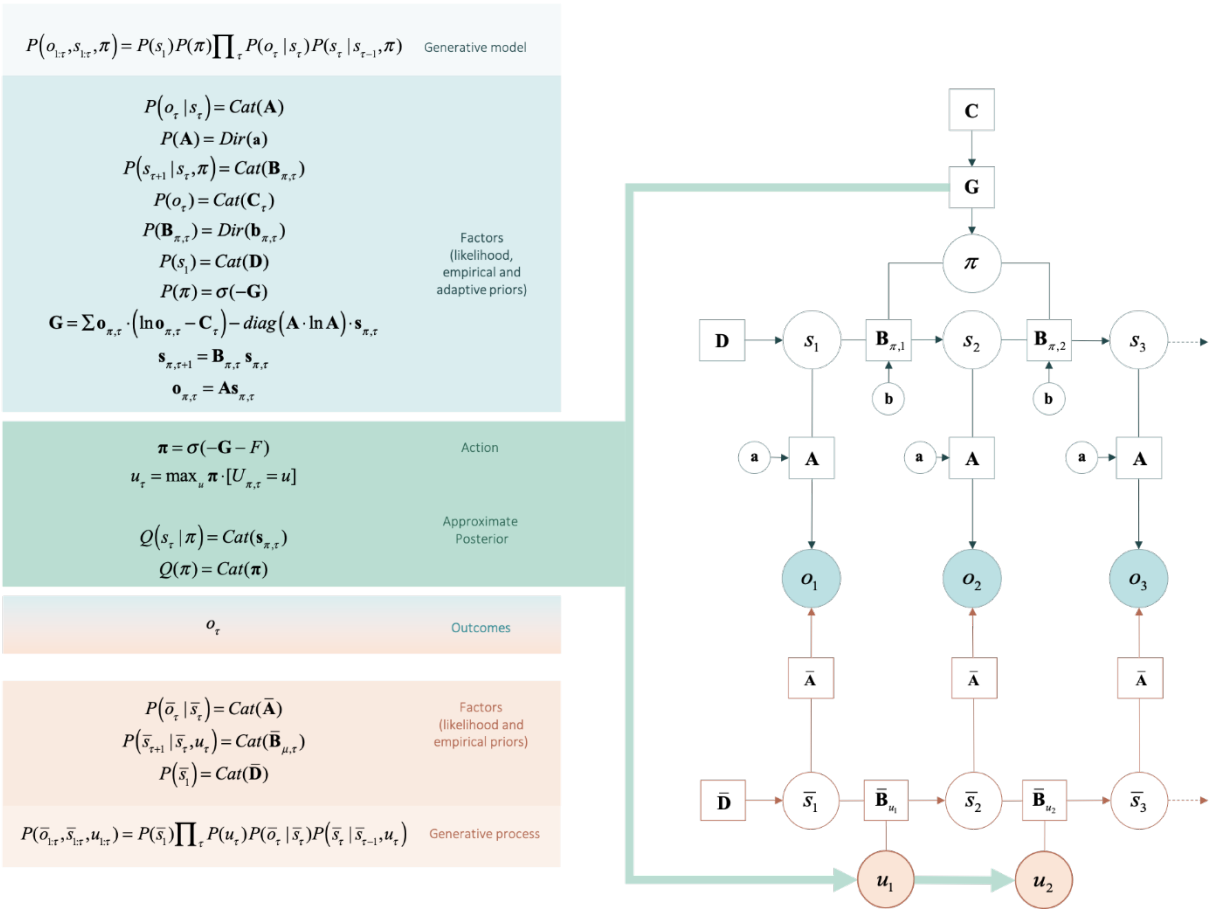


Figure 2. Computational description of the decision-making task. The generative model and generative process of our decision-making task. Open circles represent random variables (hidden states and policies), filled circles represent the outcomes, squares represent model parameters (e.g., likelihood **A**, empirical priors **B**, **D**, **G**, and the evolutionary prior **C**). The generative model is shown in the upper part of the figure, while the process generating outcomes is shown in the lower part. The generative model and process are coupled through the same outcomes (*o*) and actions (*u*), where outcomes are used to infer hidden states and policies – and action is sampled from policies to change the states that are being inferred. States of the generative model are denoted by 's' while states of the generative process are denoted as 's_bar'. The generative model is a joint probability distribution over outcomes and hidden states, which can be decomposed into factors. Factors are conditional densities (categorical: *Cat*; or Dirichlet: *Dir*) that make up the priors and likelihood of the generative model. Priors that depend on random variables, such as hidden states and policies, are empirical priors (e.g., priors that are learned at a given hierarchical level or time scale). Priors that do not vary on this time scale are initialised as evolutionary priors (e.g., **C**). These are log preference vectors that rank the desirability of associated outcomes. Lower-case **a** and **b** correspond to matrices of concentration parameters for **A** and **B** respectively. The process whereby outcomes are generated decomposes into a series of belief updates: (i) Policy selection: the sequence of actions (i.e., plan or policy) is inferred under prior beliefs that the most likely policy minimizes expected free-energy (**G**); (ii) Inference about future states depends on state transitions encoded by the transition matrix (**B**) and the likelihood (**A**); (iii) Inference about outcome: the policy – with respect to the probability transitions – generates probabilistic outcomes at each time point. The likelihood of each outcome is encoded in the likelihood matrix (**A**), which attributes the probability of each possible outcome to each possible state; and (iv) Action: the agent selects the most likely action under posterior beliefs about policies. The green arrow highlights the circular causality that results when the generative model and process are coupled through outcomes and ensuing action. The process generating outcomes triggers the message-passing, under the generative model, which entails the evaluation of a policy,

from which actions are selected. Actions change states in the generative process and a new outcome is generated. Thus, the cycle of perception and action continues. Learning corresponds to updating the concentration parameters that underwrite posterior beliefs about the likelihood of the sensory matrix (**A**). Each exchange with the environment is accumulated by concentration parameters. This accumulation encodes the probability of outcomes, given hidden states – enabling the agent to learn about environmental contingencies (and the social environment to change in response to the agent's actions). The generative model and process can be defined for any scenario. The icons in the upper panel refer to changes in the generative model induced by (simulated) pharmacotherapy, or by changes in the generative process afforded by social adversity and support. These changes are described in the next figure. For a detailed description of the update equations and underlying theory, see (Friston, Parr, et al., 2017).

The generative model and process used to simulate social inference – and ensuing changes in depressed mood – are described formally in figure 2. In brief, this setup considers 5 (observable) outcomes: an outcome that sets the scene for a social choice (e.g., being at home), three levels of social reward (low, moderate and high), and an epistemic cue that reports the current context (this is a 'Go' or 'no-Go' social context) that determines Caroline's availability.

Outcomes are generated by two kinds of external states called hidden factors. The first is the context with the two levels pertaining to Caroline's availability. These hidden states are not under the agent's control. Conversely, transitions among the states of the second factor reflect the agent's choice or policy, with four levels; i.e., home, Rudolph, Caroline, social media. The two factors interact to generate outcomes. Specifically, the context (Caroline's availability) determines whether the social media state generates an (epistemic) outcome that is 'go' or 'no-go'. Put simply, this means the agent can choose to find out whether Caroline is available or not—or contact her directly—or not. The context alternates every other day, meaning that the context-sensitive outcome available to the agent changes every other day.

Given some observations, the agent can predict outcomes under a set of plans or policies, given her beliefs about (policy-dependent) transitions among different states. The policies are: (1) home to home; (2), home to Rudolf; (3) home to Caroline; (4) home to social media; (5) Rudolf to Rudolf; (6) Caroline to Caroline; (7) social media to home; (8) social media to Rudolph; (9) social medial to Caroline; (10) social media to social media.
This enables her to evaluate the expected free-energy of each policy – and use the expected free-energy as prior beliefs to form posterior beliefs, given what she has already observed. An action is generated by selecting the most likely action from the resulting posterior. And so, the cycle of perception and action continues. Notice that the coupling between the agent and the world is mediated by observable outcomes and action. The interventions corresponding to the conditions above can be modelled, either by changing the prior beliefs of the agent (about initial conditions, likelihoods or state transitions), or by changing the prosocial world in a way that responds to her choices.

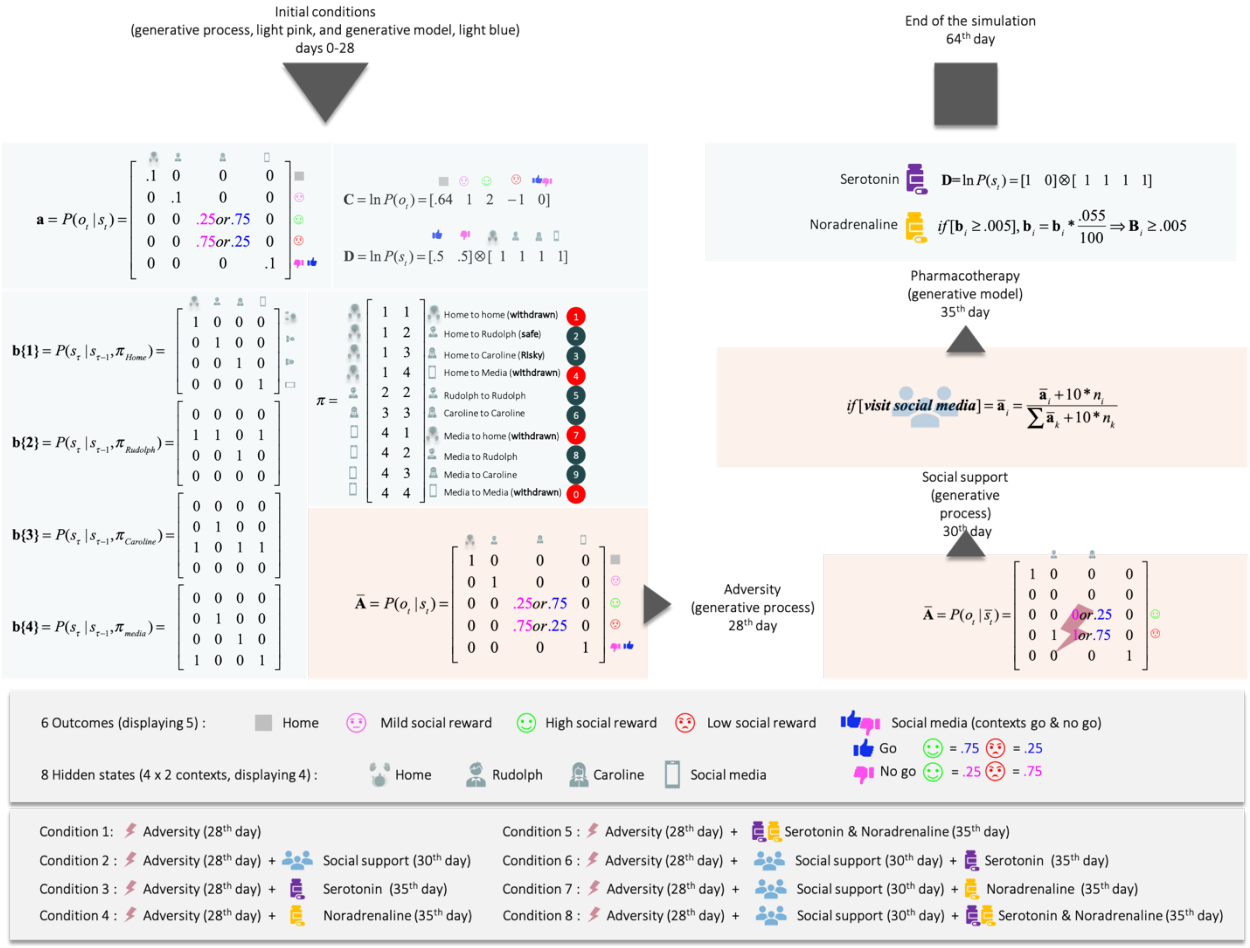


Figure 3: This figure details the likelihood and prior transition probabilities for our generative model of prosocial exchanges. The variables pertaining to the generative model are shown in light blue boxes, while the corresponding parameters of the generative process (i.e., the social world) are shown in light pink. The states and outcomes in this model are generated under two contexts pertaining to Caroline's availability: available or not available. For ease of visualization, we have shown context-sensitive outcome likelihoods. In other words, there are six potential outcomes, but we have conditioned the epistemic ('go' and 'no-go') outcome on the context (to generate five outcomes). This simplifies the graphics and is licensed by the fact that only the epistemic outcome is context-sensitive. The top-left section corresponds to the contingencies during the initial exchanges (days 0-28) and corresponds with the narrative description in Figure 1. The adverse life event on the 28th day amounts to Rudolph and Caroline (on a good day) now yielding negative outcomes, and Caroline, even on a good day, affording negative outcomes. Adversity happens when the agent is sensitive to (i.e., prone to learn) the social environment. We implemented this by reinitializing the counts over the sensory prior beliefs of the agent (**a**). Social adversity and support are modelled by changing the precision or reliability of social outcomes in the generative process – in response to social signals. This is a subtle aspect of this model; namely, the generative process or social environment responds adaptively to the agent's behaviour. As of the 30th day (for the conditions involving social support), we implement social support by adding counts (+10) to the likelihood of the environment counts (+10) for the cells corresponding to the mappings 'Rudolph and positive outcomes', 'Caroline good day context and positive outcome', and 'Caroline busy day context and negative outcome'. The n_i corresponds to the number of times the agent visited the location a_i. The increase in counts has the ultimate consequence of driving the probability mapping in the (**A**) of the generative process towards and beyond their initial values more. A '+10' is added to the cells every time the agent solicits the epistemic cue (i.e., social media). This implements the social signalling characteristic of adaptive low mood. Pharmacological interventions on the 35th day include the following: Serotonin provides an optimistic

bias by changing prior beliefs about the initial states, in favour of the 'Go' context (from .5;.5 to .99;.01). Noradrenaline decreases the precision of the transition probability matrices **B** (i.e., it increases uncertainty about future states), which leads to a gradual accumulation of uncertainty about unvisited states. Through the expected ambiguity component of expected free-energy **G**, it tends to motivate exploratory behaviours. The agent continues to learn the state transition after we administer noradrenaline.

3 Results

We used belief updating to simulate perception, action, and learning under different levels of social adversity, support and antidepressant treatment (i.e., pharmacologically induced changes in prior beliefs about states and contingencies). The results of these simulations are summarized in Figures 4-6. Behavioural outcomes and choices were assessed using the criteria listed in Table 2. In what follows, we described the responses to different scenarios or conditions in turn.

Table 2. Synthetic diagnostic criteria

Symptoms of normative depression	**Anhedonia**	**Intensity**	When the expected utility or reward is below the 95% confidence interval of the (healthy) control condition. Our subject experiences a lack of pleasure and disinterest in (prosocial) activities – of an intensity that a healthy phenotype experiences only about once every 20 days.
		Duration	When the intensity criterion is met for multiple consecutive trials. Narratively, the subject experiences a lack of pleasure and disinterest in (social) activities – lasting many days.
	Social withdrawal	Policies that do not lead to an encounter with social partners (see Figure3) 1: Stay home, stay home (starting point, Figure 1) 4: Stay home, go to social media 7: Go to social media, go back home 10: Go to social media, stay on social media	

Table 3. This table provides a summary of results in terms of the percentages of days post-adversity (out of 36) during which the synthetic subject met the subjective criteria for anhedonia in terms of intensity (expected utility below the threshold) and duration (two or more consecutive days) and the behavioural criteria for social withdrawal.

	Anhedonia		**Social withdrawal**
	Intensity criterion: expected utility below 95% CI (% of 36 days post-adversity*)	Duration criterion: 2 or more consecutive days (% of 36 days post-adversity*)	Behaviour criterium: selected policy 1,4,7, or 1 (% of 36 days post-adversity*)
CONDITION 1 Severe depression	100%	97%	61%
CONDITION 2 Adaptive mood (social support)	6%	3%	0%
CONDITION 3 Serotonin	19%	19%	81%
CONDITION 4 Noradrenaline	53%	53%	75%
CONDITION 5 Serotonin and noradrenaline combined	61%	61%	50%
CONDITION 6 Social support and serotonin combined	6%	0%	0%
CONDITION 7 Social support and noradrenaline combined	44%	31%	22%
CONDITION 8 Social support and serotonin and noradrenaline combined	6%	25%	47%

*Starting the first day after the adverse event (day 29th)

3.1 Baseline

0 Baseline (figure 4): The first 28 trials are equivalent across all simulations. In the absence of adversity, the agent skilfully responds to contextual changes by shifting between action policies that yield a (risky) high social reward and a (safe) moderate social reward. After the 7th trial, the agent always engages epistemic policies; foraging on social media first, followed by exploitative behaviour resulting in positive or negative outcomes. Before the 7th trial, the agent is still learning her prior beliefs about social partners and figuring out what policy will best suit her preferences; hence the different policies (6,9,8,4, see figure 4, bottom right for a visual description of each policy). After the 23rd day, the agent misreads the situation: Caroline was having a good day, but the agent perceived a negative outcome (e.g., by misinterpreting Caroline's behaviour during the encounter). The expected utility remains high overall (above the baseline; the pink line), and crucially, there are no consecutive days of anhedonia. Figures 5 and 6 use the same format as the upper panel in Figure 4 to show the effects of various interventions on social adversity and support, with or without pharmacological interventions.

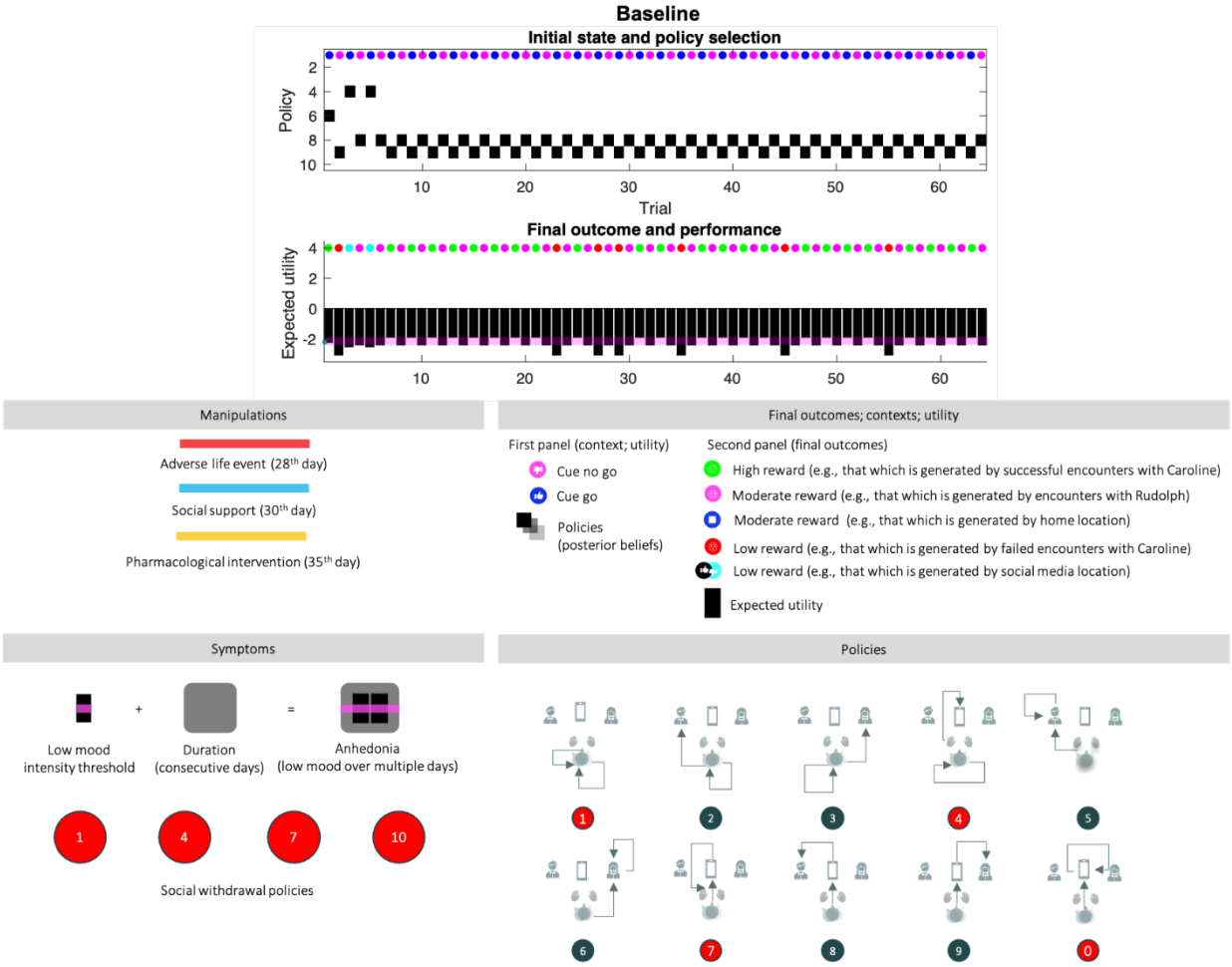


Figure 4. Baseline. **Top panel**: The upper images show the posterior expectation of each of 10 policies (see method, Figure 2) as they evolve from day to day (64 in total). The small circles in the upper part of these panels indicate the observed outcomes (context in the first panel, and outcomes in the second). The context changes every other day. The pragmatic value of these outcomes is shown as a (black) bar chart in the second panel.

The lower panel describes the interventions that depend on the condition, and the symptoms, which are: (i) anhedonia when pragmatic value or reward (black bars) is below the pink bar over multiple days (duration, black shaded rounded rectangles), and (ii) social withdrawal, expressed by policies 1,4,7, and 10. The lower left panel provides a legend (upper) and a graphical description of the policies (lower). The intensity component of anhedonia corresponds inversely to the expected utility of a policy, or the extent to which it will yield preferred outcomes. Narratively speaking, this amounts to expecting socially rewarding outcomes when engaging a certain action. The intensity component of anhedonia is thus defined as low appetitive action. We assume that normal levels of appetitive action correspond to the expected utility experienced on most days, for a healthy (baseline) agent (pink line). The duration component of anhedonia corresponds to the number of consecutive days. A normative assessment of anhedonia thus would involve 14 consecutive days, as is the case in the condition of severe depression below.

3.2 Severe depression, social support, serotonin and noradrenaline

1 Severe depression (figure 5, upper left quadrant): The agent experiences social adversity on the 28th trial (i.e., a rejection from Rudolph and Caroline), and has no social support (i.e., her signalling has no effect on Caroline and Rudolph). The adverse life event entails ongoing exposure to negative outcomes. The increase in exposure to negative outcomes is caused by a change in the generative process, which now yields 0% chance of generating a mildly rewarding outcome at the Rudolph state (previously 100% chance), and a 100% chance of generating a negative outcome at that same state. In addition, there is now a 0% chance of a positive outcome and a 100% chance of a negative outcome at the Caroline state during the no-go context (busy day), and the probability of Caroline yielding a positive outcome on a good day has been inverted. Now, even on a good day, Caroline only affords a 25% chance of a positive outcome (see adversity on the 28th day, fig. 3). Importantly, the adverse life event affects both the generative process (making bad outcomes more likely for Caroline, and unavoidable for Rudolph), and the generative model by resetting the concentration parameters to their initial values (as they were at trial 1). The motivation for reinitializing the counts is primarily to sensitize our agent to novel outcomes. This sensitization rests on the fact that learning slows down with the accumulation of concentration parameters (e.g., during the first 27 days). Because the agent's generative model reverts to its initial settings, the agent expects to obtain positive outcomes at Caroline's on her good days, for some time after the adverse life event. This explains why our agent keeps selecting policy 9, which leads to Caroline, on multiple days (8 days) after the adverse event.

In our simulation, such a manipulation does not map onto a biological process that we would have aimed to reproduce in silico. It is simply an artefact of the design. Narratively, it may be said that it simulates the awareness of a change in social context caused by a functional forgetting in short-term memory (i.e., from trial 1 through 28) reinstating the agents initial memory parameters (i.e., at trial 1). This leads to an increased sensitivity to the novel social environment. In this particular sense, the resetting of concentration parameters is arguably consistent with the phenotype of depression. Early childhood adversity is a risk factor for depressive disorder by sensitizing the individual to proximal environmental stressors later in life —e.g., making the agent more likely to undergo parameter reset after an adverse life event (Starr et al., 2014) and memory disruptions and negative biases are commonly associated with depression—e.g., acquiring negative bias

based on the learning of pessimistic expectations after adverse life events (Dillon & Pizzagalli, 2018). A simulation explicitly aimed at studying the impact of functional forgetting on treatment course could either systematically vary the depth of forgetting or use hierarchical models to allow forgetting to emerge naturally from learning and inference of higher-level contextual states (Hesp et al., 2021).

Occasionally, the agent experiences a negative outcome when Caroline was supposedly having a good day (as indicated by the Go cue). This occurs on average about 25% of the time, because outcomes are generated from the likelihood mapping in Figure 3, which shows there are intrinsic uncertainties in the mapping from Caroline's mood to positive or negative outcomes (25% chance of failure on Caroline's good days, 75% chance of failure on Caroline's bad days). On average, the agent will get a dissatisfying outcome 25% of the time the agent visits Caroline on a good day, because of the constitution of the likelihood mapping (figure 3). The agent is not misinterpreting the cue. It is Caroline that exhibits intrinsic variability.

The agent persistently evinces a low mood, below baseline (i.e., intensity of anhedonia). 14 days after the adverse live event, the agent shifts to a social withdrawal policy (4). This is caused by acquiring a pessimistic likelihood about the outcomes afforded by Caroline and Rudolph. Without intervention, the pessimistic likelihood is successively reinforced.

2 Social support (figure 5, upper right quadrant): In this scenario, the agent experiences adversity on day 28 but is provided with social support 2 days later (i.e., her social signalling changes Caroline's and Rudolph's behaviour). Following this, the agent's mood recovers, relative to the baseline condition. This is because Caroline becomes more reliable and the agent is certain that Caroline will show up on a good day, and not on a busy day. This scenario corresponds to what is expected under both the social risk hypothesis and our EST of depression. When the environment is adaptive (i.e., responsive), low mood causes the agent to regain typical functioning – via social signalling. Note that social support failed in simulations where the support was delayed by more than 2 days. After 2 days without support, the pessimistic beliefs become too robust, and no amount of social support is enough to reshape the prior. When the support comes too late, the agent spirals into severe depression. Of course, the critical period of intervention of 2 days depends on the parametrisation of the generative model. Under different parameter values, the critical period could be extended. This speaks to the importance of the timing of social interventions to effectively interrupt and revert the learning of the pessimistic likelihood. More formally, the adaptive response comes from a change in the likelihood of the generative process (see fig. 3), which by generating certain outcomes, leads to a learning of the likelihood matrix. This learning assigns high probabilities to the mappings between the Rudolph state and the high social reward (instead of the moderate social reward), between the Caroline state and the high social reward on the 'go' context, and between the Caroline state and the low social reward on the 'no go' context. The behavioural manifestation of the social intervention is a return to the correct policy, given the context, namely 8 and 9. We now consider the therapeutic effects of pharmacotherapy in the absence of social support.

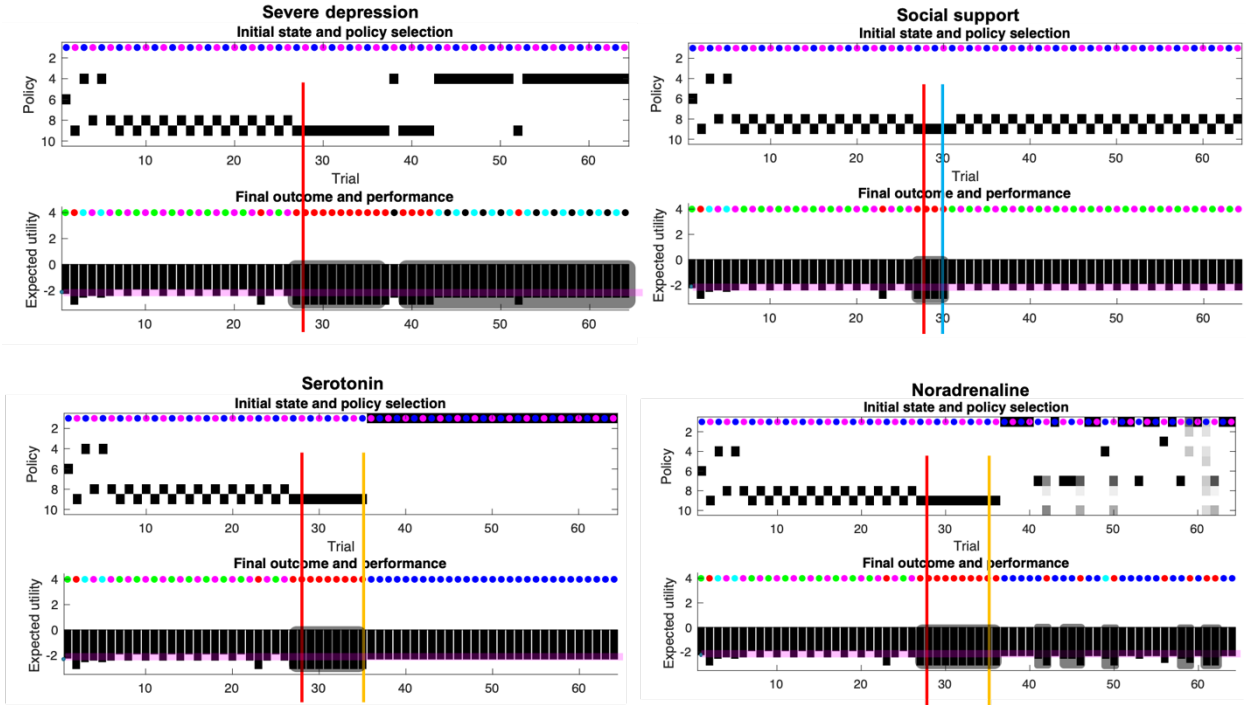
3 Serotonin (figure 5, lower left quadrant):

Serotonin upregulates prior expectations over the 'go' state at the beginning of the trial (**D**). The intervention based solely on serotonin precludes consecutive days of low mood. However, social withdrawal remains

(policy 1). Given that the agent receives no social support, the likelihood of receiving negative outcomes from Rudolph on either day is still 100%. The likelihood remains pessimistic after the social adversity on the 28th day; hence the best move for our agent is to stay at home (policy 1), despite the serotonergic bias on beliefs over the go context.

Administration of serotonergic antidepressants induces very strong expectations of Caroline having a 'good day', which had the (unintended) side-effect of countering our agent's epistemic drive. The agent experiences multiple bad outcomes between the moment of the adverse life event and the beginning of the pharmacotherapy, even on good days. The consequence of this is that the expected utility of good days reduces as the agent is left with neither an epistemic nor a pragmatic drive—and opts to stay at home instead. This slightly counter-intuitive effect of serotonergic pharmacotherapy underscores the clinical relevance of (1) the timely administration of antidepressants (e.g., before further negative associations become dominant), (2) the support of antidepressants with other types of interventions (i.e., this effect does not occur when combined with social support in our simulations), and (3) the further investigation of potential ways to model and predict the (side-)effects of antidepressants.

4 Noradrenaline (figure 5, lower right quadrant): Noradrenaline gradually increases uncertainty about future states (i.e., increases uncertainty in the transition **B** matrices), which underwrites a loss of precise belief-updating during planning—and motivates exploratory behaviours, through the expected ambiguity (in **G**). This is reflected in figure 5 (lower right quadrant, top panel) showing imprecise beliefs over policies 1 to 10. Noradrenaline intervention engenders several days of low mood after administration, which are generally associated with social withdrawal (policy 7). There is a combination of withdrawal policies (1,4,7), and uncertainty over these policies; e.g., the agent sometimes ends up going to Caroline and receiving a negative outcome (e.g., day 42). Episodes of anhedonia and social withdrawal are short, but present nonetheless, which suggests that the agent is still depressed. We next turn to the effects of combining pharmacotherapy with social support by repeating the above conditions in the setting of a responsive social context.



Figures 5. Responses to intervention. This figure uses the same format as the upper panels of Figure 4. Interventions are indicated by the solid lines (red line: social adversity; blue line: social support; orange line: pharmacotherapy). The plots report the simulated responses to social adversity (red lines in all quadrants), and the remedial effects of social support (blue line in the second quadrant). Quadrants with orange lines show the corresponding effects of pharmacotherapy (serotonin or noradrenergic). The four treatment conditions show the same behaviour over the first 28 days as the baseline scenario. This figure reports the results of conditions 1 to 4.

3.3 Combined interventions

Computationally, social support, serotonin, and noradrenaline operate the same way as described above, whether they are administered individually or combined. What changes are the behavioural effects. To understand these novel effects, we must pay attention to the temporal structure of the depressed system (i.e., the coupled generative model and process). Social support will be the first intervention to impact the generative model (the agent part of the system) by generating the outcome on the basis of which inference and learning operate. Serotonin will act first by influencing initial states (**D**), and finally noradrenaline will act by influencing policy planning (through **B**). The selected policy, if it involves going to the social media state, will influence the probability of outcomes in the generative process (see 30th day, fig. 3), which will then loop back into the generative model to influence inference and learning.

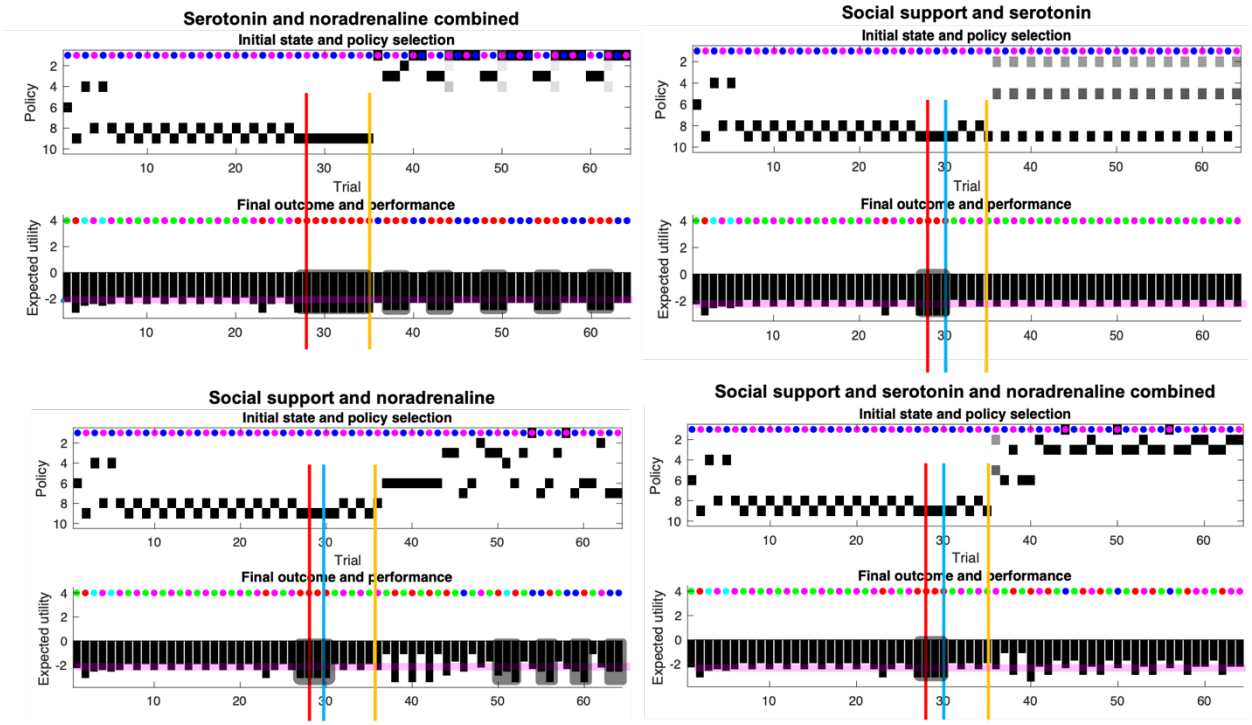
5 Serotonin and noradrenaline combined (figure 6 lower left quadrant): After the pharmacotherapy on day 35, the agent experiences episodes of anhedonia at regular intervals. However, these are characterized by perceived negative social encounters with Caroline, not social withdrawal. According to the specifications of our simulation, this means that the agent does not meet the requirements for severe depression (i.e., anhedonia and withdrawal criteria). The agent alternates between policies 1, 2, and 3, which do not involve

going to social media. This is arguably because serotonin promotes an optimistic bias, meaning that no information foraging is required (e.g., going on social media).

6 Social support and serotonin (figure 6, lower left quadrant): In this condition, there is no withdrawal and overall, the mood states are non-depressed (above baseline). This condition combines an optimistic bias with an increase in social stability, yielding high certainty about the reception the agent will receive from Caroline and Rudolph. Since beliefs about policy-dependent state transitions remain the same, there is no need to explore. On Caroline's good days, the agent approaches Caroline, and on her busy days, the agent engages Rudolph. Note, however, that the agent remains uncertain about which policy to pursue, and compared to the scenario combining noradrenaline, serotonin, and social support, the agent never engages pragmatic policies (e.g., 6).

7 Social support and noradrenaline (figure 6, lower right quadrant): This condition yields a variety of responses, and some short episodes of low mood. These are sometimes caused by social withdrawal (e.g., days 64,63), and sometimes by high risk-taking (e.g., day 60), expressed by policy 6. The exploration of the policy space in this scenario is driven by the slow decrease in precision over the transitions (**B** matrices), coupled with an increase in social partners' reliability.

8 Social support and serotonin and noradrenaline combined (figure 6, upper right quadrant): The agent experiences adversity but has social support and access to pharmacotherapy. This scenario largely precludes social withdrawal and consecutive days of anhedonia, and the agent is highly optimistic. Almost on every occasion, the agent engages policy 3 (i.e., wait, then approach Caroline), which explains mood episodes below baseline. Otherwise, the agent engages policy 2 (i.e., wait, then approach Rudolph). Low mood is characterized by risk taking, not social withdrawal. Moreover, for the first time, the agent engages policy 6, which is a pragmatic policy (i.e., going directly to Caroline). This speaks to the effect of noradrenaline, which motivates the agent to disambiguate (future) states that are deemed uncertain, while the few days of withdrawal speaks to the serotonergic bias manifest when Caroline is on a bad day. Note that the epistemic character of a policy concerns the extent to which it disambiguates uncertain transitions. Now, uncertain transitions might be transitions between non epistemic 'states', that is, states that provide go/no-go outcomes (i.e., if I know where the cue is and where the cue leads, but I do not know if my current location leads to a reward, I will explore this latter transition first, especially if I believe I am in a 'go' context, which is what the serotonergic bias does). Hence this condition involves epistemic policies—as in disambiguating behaviour—without these policies soliciting epistemic cues.



Figures 6. Responses to pharmacotherapy and social support. This figure uses the same format as Figure 4 and 5. Here, we report the responses to the final three conditions; namely, responses to serotonin and noradrenaline and combinations of drug treatment (yellow line), after social support (blue line).

4 Discussion

Using active inference, we have reproduced (artificial) anhedonia and social withdrawal to provide a numerical analysis of the EST of depression. We specified a generative model, involving multiple components that conspired to generate context-sensitive responses to social uncertainty; particularly, the prior preferences for socially rewarding outcomes (e.g., encounters with Caroline). Our results provide support for our hypothesis that depressed mood reflects an adaptive response to interpersonal adversity. Following an adverse life event, our synthetic agent resolved interpersonal uncertainty via social signalling, thereby alleviating her depressed mood. Except for scenarios involving social support, all the conditions we simulated resulted in an above-average duration of episodes of anhedonia and social withdrawal, speaking to unresolved uncertainty. Crucially, we do not claim that depressive psychopathology is adaptive. Indeed, unlike our 'social support' condition, the 'severe depression' scenario proved to be maladaptive, characterized by unresolvable episodes of low mood and social withdrawal. This may either occur when signalling is defective (e.g., due to personality difficulties, rendering a person unable to deliver the appropriate signals), or when it fails to be received (e.g., cues provided by someone who is socially isolated).

A key aspect of simulations similar to ours is the explicit and formal modelling of the aetiological factors that underwrite the selection of, or inference about, prosocial behaviour; ranging from Bayesian belief updating (i.e., perceptual inference), through to experience-dependent plasticity (i.e., perceptual learning), and to the social and encultured responses of the environment. The active inference framework has an explicit (neuronal)

process theory, which could allow future studies to simulate the selective effects of neuromodulatory interventions on the encoding of precision or uncertainty, and its consequences for the agent's social behaviour. Having a complete model of (aberrant) social inference means that in the future, one could simulate neuronal processes that lend themselves to empirical measurement. Studies along these lines could simulate dopamine responses in order to provide qualitative predictions that could be tested with functional magnetic resonance imaging, e.g.,(D'Ardenne et al., 2008; Schwartenbeck et al., 2015). In this setting, dopamine responses are usually associated with updates to the expected precision of Bayesian beliefs about the policy in play (see Appendix E in Friston, FitzGerald, et al., 2017; Sales et al., 2019)

In our simulations, we allowed uncertainty over contingencies to decrease every time the agent referred to social media. Narratively, this could be interpreted as the agent signalling (implicitly or explicitly) to Caroline and Rudolph that they should be more consistent in order to provide more support (e.g., by manifesting discontent or by sharing emotional state on her social media feed). Note that this does not imply any qualitative change in prosocial responses; it simply corresponds to an increase in the consistency or reliability of responses that may or may not be affiliative. Computationally, this amounts to repairing the environment, such that the prior beliefs of a phenotype are fit for purpose. In other words, the (social) environment changes to match the prior beliefs of its incumbents; thereby reversing the suboptimality implicit in maladaptive depression. Recovery then depends on the sensitivity of the subject's social environment and on how often she consults social media. Here, the positive effect of combining pharmacotherapy with social support is thought to be attributable to the optimistic bias associated with serotonin (Harmer, 2008; Harmer et al., 2017), coupled with the effect of noradrenaline, which motivates the exploration of states associated with rewards (Aston-Jones & Cohen, 2005).

Given the parametrization of our subject, the best intervention was the combination of social support and serotonin, while the worst outcome—in terms of social withdrawal—was the intervention with serotonin alone (see results table). How these two interventions work together in real participants remains open to question. Computationally, serotonin provides an optimism bias while social support confirms that bias by returning the social environment to its normal setting (i.e., the setting matches the non-pessimistic expectations of the agent). However, when social support is lacking, serotonin leads to repeated, failed social encounters and social isolation. It is unclear whether serotonergic antidepressants are direct mood enhancers. Rather, it is suggested that antidepressants work by augmenting positive emotional processing, which then has positive effects on other psychological factors (Harmer et al., 2009). Our simulation results highlight this more complex systemic interaction between the psychosocial and neurocognitive aspects of depression and stresses the importance of social support. Indeed, social support in older adults is known to have alleviating, bidirectional effects on symptoms of depression and anxiety. Social disconnectedness appears to predict perceived isolation, which itself predicts higher depressive symptoms, and *vice versa* (Santini et al., 2020). Adolescents who self-report higher perceived social support at age 19 are less likely to show depressive symptoms one year later (Scardera et al., 2020), and reviews emphasize the significant protective effects of perceived emotional and instrumental support, as well as social network diversity in the general population (Santini et al., 2015). The strength of the positive effect of social support, of course, rests on the subject-specific parametrization, which we can expect to vary across real subjects. For instance, we initialized our subjects as 'blank slates' with respect to state transitions. However, this would be expected to vary across participants based on their

individual experiences and development. This may also vary based on the volatility of the (prosocial) environment prior to the occurrence of social adversity. Again, our proposal is a proof of principle, and is only meant as a general portrait of what is feasible, when considering the social environment in computational phenotyping.

Our results speak to the Darwinian models of depression synthesized by the Social Risk Hypothesis (SRH) (cf. box 1). Following the attachment model, simulated agents – displaying anhedonia and social withdrawal – inhibited social risk-taking under social uncertainty. Following the social competition hypothesis, the adverse life event reduced social uncertainty by producing social withdrawal. Consistent with the resource conservation model, after the adverse event, the agent progressively returned to Caroline, so long as the agent knew exactly when to approach her. From the point of view of the SRH, the explanations of the attachment model, the social competition model, and the resource conservation model are all grounded in the dynamics we simulated. The dynamics we simulated were the increase in social uncertainty leading to behavioural and psychological symptoms that either lead to depression—when the social environment is not responsive—or to the restabilization of the social network—when the social environment is responsive. The Evolutionary System Theory (EST) of depression, which is the recent neurocomputational reinterpretation of the SRH, frames the adaptive mood dynamics integrated by the SRH as an attunement dynamic between evolutionary adaptive priors (here prior preferences), plastic developmental priors (here B and a likelihood A), and a social environment (here a generative process). These would have been selected to conspire to generate adaptive symptoms of depression in order to trigger social network re-stabilization (ex. condition 2); social network stability having been crucial to evolutionary success throughout human history (see box 1). When the social environment fails to respond to the social signalling represented by depressive symptoms, the behaviourally adaptive pessimistic beliefs that produced this signalling spirals into the maladaptive beliefs, characteristic of depressive illness.

5 Conclusion: future directions

Simulation studies such as ours can be used to simulate both the symptoms and underlying processes of inference *in silico*. Note, however, that our generative model only had one level. By adding levels to the generative model, as in hierarchical (deep) active inference (Friston, Parr, et al., 2017), one could further fine-tune these affective dynamics. For instance, one could keep lower-level preferences fixed, reflecting their evolutionary origins, while allowing learning in higher-level preferences to change as a function of life experiences (e.g., learning to prefer Rudolph's underwhelming calm over Caroline's extravagance). Furthermore, generative models — of the kind used above — can be fitted to individual and population level clinical data; involving some general-purpose tasks related to a disorder of interest (e.g., social decision-making in depression), thereby yielding a novel avenue for computational phenotyping, prognosis, and diagnostic nosology. The idea here is that clinicians could then predict psychiatric trajectories in specific individuals, when conditioned on different available treatment options. The latter could then be used to generate a prognosis and course of treatment tailored for any client, which we believe is perhaps the most exciting promise of generative modelling in clinical psychiatry.

However, before achieving this, there are many conceptual and technical limitations to overcome, which chiefly relate to the treatment of clinical data using computationally meaningful generative models. Behavioural measurements such as hits and misses and associated social withdrawal can be measured in experimental designs that track behaviour in a decision-making task, with a given narrative (e.g., based on vignettes of real-life scenarios). The challenge lies in fitting individual and environmental initial conditions for both the generative model and generative process (see method, Figure 3). For instance, assuming that preferences are endowed by (encultured) evolution, one should provide a reliable estimate of population-level preferences for social encounters. Then, one should assess the degree of precision of empirical priors and measure the expected utility for each action policy. Crucially, in order to implement the effect of social support, one could also gather and translate information about environmental responsivity. This could be done via task-specific questionnaires (e.g., on a Likert scale, how desirable is an encounter with Rudolph versus Caroline? How reliable do you consider Caroline? Etc.). Alternatively, these questions could be answered by data captured by various technologies. Smartphone-based, passive sensing technologies, which can capture behavioural data (e.g., distances travelled, exercise, sleep, social media activity) and psychological data (e.g., affective tone of text entered), might help in this regard (Sapiro et al., 2019). More generally, the specification of environmental components might be achieved by using various local cultural factors (e.g., cultural norms) regarding the responsivity to idioms of distress; i.e., culturally specific ways of expressing illness experience (Kirmayer & Young, 1998).

In short, to achieve clinical utility, generative models of depression should summarize the client's neurocognitive disposition to learning as well as her social situation, in terms of the environmental responsivity to her signalling. The role of the clinician, then, would be to map the evolutionary (e.g., adaptive priors), neurocognitive (e.g., empirical priors), and social (e.g., environmental responsivity) portrait of specific clients in terms of a generative (phenotypic) model – a Computational Evolutionary Social assessment of sorts. This opens a novel avenue for research, which attempts to quantify both generative models and processes, by bringing together the expertise of cultural, evolutionary, and computational psychiatrists and psychologists. If such an approach proves reliable – and robust predictions can be made regarding the course of illness experience and optimal treatment options – using computational (social and neurocognitive) phenotyping to improve psychiatric assessment, diagnosis, and tailored interventions might become commonplace.

References

Allen, N. B., & Badcock, P. B. T. (2003). The social risk hypothesis of depressed mood: evolutionary, psychosocial, and neurobiological perspectives. *Psychological Bulletin*, *129*(6), 887–913.

Allen, N. B., & Badcock, P. B. T. (2006). Darwinian models of depression: a review of evolutionary accounts of mood and mood disorders. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, *30*(5), 815–826.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. American Psychiatric Pub.

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annual Review of Neuroscience*, *28*, 403–450.

Badcock, P. B., Davey, C. G., Whittle, S., Allen, N. B., & Friston, K. J. (2017). The Depressed Brain: An Evolutionary Systems Theory. *Trends in Cognitive Sciences*, *21*(3), 182–194.

Badcock, P. B., Friston, K. J., & Ramstead, M. J. D. (2019). The hierarchically mechanistic mind: A free-energy formulation of the human psyche. *Physics of Life Reviews*. https://doi.org/10.1016/j.plrev.2018.10.002

Badcock, P. B., Friston, K. J., Ramstead, M. J. D., Ploeger, A., & Hohwy, J. (2019). The hierarchically mechanistic mind: an evolutionary systems theory of the human brain, cognition, and behaviour. *Cognitive, Affective & Behavioral Neuroscience*. https://doi.org/10.3758/s13415-019-00721-3

Bruineberg, J., Rietveld, E., Parr, T., van Maanen, L., & Friston, K. J. (2018). Free-energy minimization in joint agent-environment systems: a niche construction perspective. *Journal of Theoretical Biology*. https://doi.org/10.1016/j.jtbi.2018.07.002

Buckner, J. D., Joiner, T. E., Jr, Pettit, J. W., Lewinsohn, P. M., & Schmidt, N. B. (2008). Implications of the DSM's emphasis on sadness and anhedonia in major depressive disorder. *Psychiatry Research*, *159*(1-2), 25–30.

Constant, A., Ramstead, M. J. D., Veissière, S. P. L., Campbell, J. O., & Friston, K. J. (2018). A variational approach to niche construction. *Journal of the Royal Society, Interface / the Royal Society*, *15*(141). https://doi.org/10.1098/rsif.2017.0685

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.

Cullen, M., Davey, B., Friston, K. J., & Moran, R. J. (2018). active inference in OpenAI Gym: A Paradigm for Computational Investigations Into Psychiatric Illness. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging*, *3*(9), 809–818.

D'Ardenne, K., McClure, S. M., Nystrom, L. E., & Cohen, J. D. (2008). BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. *Science*, *319*(5867), 1264–1267.

Dillon, D. G., & Pizzagalli, D. A. (2018). Mechanisms of Memory Disruption in Depression. *Trends in Neurosciences*, *41*(3), 137–149.

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nature Reviews. Neuroscience*, *11*(2), 127–138.

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O Doherty, J., & Pezzulo, G. (2016). active inference and learning. *Neuroscience and Biobehavioral Reviews*, *68*, 862–879.

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: a process theory. *Neural Computation*, *29*(1), 1–49.

Friston, K. J., Parr, T., & de Vries, B. (2017). The graphical brain: Belief propagation and active inference. *Network Neuroscience*, *1*(4), 381–414.

Gilbert, P. (1997). The evolution of social attractiveness and its role in shame, humiliation, guilt and therapy. *The British Journal of Medical Psychology*, *70 (Pt 2)*, 113–147.

Harmer, C. J. (2008). Serotonin and emotional processing: does it help explain antidepressant drug action? *Neuropharmacology*, *55*(6), 1023–1028.

Harmer, C. J., Duman, R. S., & Cowen, P. J. (2017). How do antidepressants work? New perspectives for refining future treatment approaches. *The Lancet. Psychiatry*, *4*(5), 409–418.

Harmer, C. J., Goodwin, G. M., & Cowen, P. J. (2009). Why do antidepressants take so long to work? A cognitive neuropsychological model of antidepressant drug action. *The British Journal of Psychiatry: The Journal of Mental Science*, *195*(2), 102–108.

Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K. J., & Ramstead, M. J. D. (2020). Deeply Felt Affect: The Emergence of Valence in Deep Active Inference. *Neural Computation*, 1–49.

Hindash, A. H. C., & Amir, N. (2012). Negative Interpretation Bias in Individuals with Depressive Symptoms. *Cognitive Therapy and Research*, *36*(5), 502–511.

Hrdy, S. B. (2011). *Mothers and others*. Harvard University Press.

Huys, Q. J. M., Daw, N. D., & Dayan, P. (2015). Depression: a decision-theoretic analysis. *Annual Review of Neuroscience*, *38*, 1–23.

Huys, Q. J. M., Guitart-Masip, M., Dolan, R. J., & Dayan, P. (2015). Decision-Theoretic Psychiatry. *Clinical Psychological Science*, *3*(3), 400–421.

Ingram, R. E., Miranda, J., & Segal, Z. V. (1998). *Cognitive vulnerability to depression*. Guilford Press.

Kaplan, R., & Friston, K. J. (2018). Planning and navigation as active inference. *Biological Cybernetics*. https://doi.org/10.1007/s00422-018-0753-2

Kirmayer, L. J., & Young, A. (1998). Culture and somatization: clinical, epidemiological, and ethnographic perspectives. *Psychosomatic Medicine*, *60*(4), 420–430.

Kirmayer, L. J., & Young, A. (1999). Culture and context in the evolutionary concept of mental disorder. *Journal of Abnormal Psychology, 108*(3), 446–452. https://doi.org/10.1037/0021-843X.108.3.446

Klinger, E. (1975). Consequences of commitment to and disengagement from incentives. *Psychological Review*, *82*(1), 1.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, *16*(1), 72–80.

Nesse, R. M. (1990). Evolutionary explanations of emotions. *Human Nature* , *1*(3), 261–289.

Nesse, R. M. (2000). Is depression an adaptation? *Archives of General Psychiatry*, *57*(1), 14–20.

Nettle, D. (2004). Evolutionary origins of depression: a review and reformulation. *Journal of Affective Disorders*, *81*(2), 91–102.

Parr, T., & Friston, K. J. (2017). Uncertainty, epistemics and active inference. *Journal of the Royal Society, Interface / the Royal Society*, *14*(136). https://doi.org/10.1098/rsif.2017.0376

Parr, T., Rees, G., & Friston, K. J. (2018). Computational Neuropsychology and Bayesian Inference. *Frontiers in Human Neuroscience*, *12*, 61.

Price, J. (1967). THE DOMINANCE HIERARCHY AND THE EVOLUTION OF MENTAL ILLNESS. *The Lancet*, *290*(7509), 243–246.

Rude, S. S., Valdez, C. R., Odom, S., & Ebrahimi, A. (2003). Negative Cognitive Biases Predict Subsequent Depression. *Cognitive Therapy and Research*, *27*(4), 415–429.

Sales, A. C., Friston, K. J., Jones, M. W., Pickering, A. E., & Moran, R. J. (2019). Locus Coeruleus tracking of prediction errors optimises cognitive flexibility: An active inference model. *PLoS Computational Biology*, *15*(1), e1006267.

Santini, Z. I., Jose, P. E., York Cornwell, E., Koyanagi, A., Nielsen, L., Hinrichsen, C., Meilstrup, C., Madsen, K. R., & Koushede, V. (2020). Social disconnectedness, perceived isolation, and symptoms of depression and anxiety among older Americans (NSHAP): a longitudinal mediation analysis. *The Lancet. Public Health*, *5*(1), e62–e70.

Santini, Z. I., Koyanagi, A., Tyrovolas, S., Mason, C., & Haro, J. M. (2015). The association between social relationships and depression: a systematic review. *Journal of Affective Disorders*, *175*, 53–65.

Sapiro, G., Hashemi, J., & Dawson, G. (2019). Computer vision and behavioural phenotyping: an autism case study. *Current Opinion in Biomedical Engineering*, *9*(C). https://par.nsf.gov/biblio/10096115

Scardera, S., Perret, L. C., Ouellet-Morin, I., Gariépy, G., Juster, R.-P., Boivin, M., Turecki, G., Tremblay, R. E., Côté, S., & Geoffroy, M.-C. (2020). Association of Social Support During Adolescence With Depression, Anxiety, and Suicidal Ideation in Young Adults. *JAMA Network Open*, *3*(12), e2027491.

Schwartenbeck, P., FitzGerald, T. H. B., Mathys, C., Dolan, R., & Friston, K. (2015). The Dopaminergic Midbrain Encodes the Expected Certainty about Desired Outcomes. *Cerebral Cortex* , *25*(10), 3434–3445.

Schwartenbeck, P., & Friston, K. (2016). Computational Phenotyping in Psychiatry: A Worked Example. *eNeuro*, *3*(4). https://doi.org/10.1523/ENEURO.0049-16.2016

Schwartenbeck, P., Passecker, J., Hauser, T. U., FitzGerald, T. H., Kronbichler, M., & Friston, K. J. (2019). Computational mechanisms of curiosity and goal-directed exploration. *eLife*, *8*. https://doi.org/10.7554/eLife.41703

Starr, L. R., Hammen, C., Conway, C. C., Raposa, E., & Brennan, P. A. (2014). Sensitizing effect of early adversity on depressive reactions to later proximal stress: Moderation by polymorphisms in serotonin transporter and corticotropin releasing hormone receptor genes in a 20-year longitudinal study. In *Development and Psychopathology* (Vol. 26, Issue 4pt2, pp. 1241–1254). https://doi.org/10.1017/s0954579414000996

Conclusion to chapter 4

A theoretical problem that my mentors, Paul Griffiths and Paul Badcock, brought up multiple times with the simulation of chapter 4 was the lack of "developmental" parameters, or the fact that what we called "evolutionary" or "adaptive" priors should be in fact priors that are also learnable. For instance, Paul Badcock proposed that the entire model itself should be viewed as "evolutionary". At first, I could not see why this view should be favoured, and so, I resisted and suggested that the evolutionary component of the model remained this one prior preference. Now, with more perspective on the model, I think that Paul was right. It is the entire architecture of the model that should be the "evolutionary" prior (although it includes both priors and likelihood). And the generative process — the social environment — should probably be viewed as the 'cultural' counterpart of the process, driving gene-culture coevolutionary dynamics (e.g., by allowing the selection of culture-characteristic priors over multiple generations of models, whose development would be driven by environmental observations characteristic of the host generative process). A subset of priors should be "empirical", that is learnable over development, and the architecture of the model (e.g., the fact that a prior does exist or not in the model) should be viewed as an evolutionary fixation (e.g., a genetic prior). Theoretically, this would be one way to respond to the problem raised by Paul Griffith, who encouraged me to think about what a developmental prior would look like in the model of chapter 4.

Another discussion I would have liked to have, but that would have been outside of the scope of this paper, was the discussion of the relation between the practice of engineering such a multi-scale simulation and the healing practice in psychiatry. From an engineering point of view, the main challenge was to induce pathological behaviour by manipulating external parameters while being able to return the model to normal functioning via internal and/or external parameters. Because of the complex interactions between the various internal and external parameters of the model, it was a challenge to return the agent to normal mood, as I had to find the right timing and order of interventions, both social and pharmacological, each acting at various spatiotemporal scales (e.g., fast-changing neural parameters and slow changing social parameters) – note that a more clever coder would certainly have taken the time to find a way to automatize (e.g., amortize) that process.

At any rate, it is worth noting that this was not only a theoretical problem that spoke to potentially important clinical implications when working with a multiscale conceptual model. From an engineering point of view, for me, as the person running the simulation, finding the right tuning of parameters that would allow me to induce the behaviour corresponding to the synthetic depression criteria and to find the right timing for interventions was a challenge. Happily, enough, I did not have to worry too much about the consequences of my "synthetic clinical decision" with the model, as I could test the impact of various orders of interventions without having to worry about the impact on a human. Moreover, I could test all the different interventions, individually, or combined. I honestly do not know how I would have managed to find the right interventions if I had needed to consider the consequences of those on a real person.

It was brought to my attention by my supervisor Prof. Kirmayer that my experience with the modelling pointed to “a crucial challenge but also a potential benefit to modelling” (personal communication). As Prof. Kirmayer

put it, “Models are sensitive to parameter choices—and in the absence of relevant empirical studies, parameters are often chosen more or less arbitrarily. The positive implication is that, by studying the effects of parameter variation (including magnitude and timing) we can discover features of system dynamics. This could have real value in understanding why interventions do not work (as expected) and how we might intervene more effectively (e.g., with multiple, timed or staged interventions)” (personal communication). I could not agree more with Prof. Kirmayer’s view. The choice of the manipulations in the study had to be informed by the empirical literature on depression. The goal was not to claim that the intervention truly mimics the action of serotonin in the brain. Rather, the goal was to guide my decisions on what parameters to manipulate so as to move the model in one direction or the other, under the assumption that that model could capture some of the essential dynamics underwriting symptoms of depression.

This engineering challenge reflects the reason why mental disorders such as depression are so difficult to treat, and perhaps why their recurrence rate is so high (Burcusa & Iacono, 2007); because mental disorders are multiscale moving entities whose parts can change both quickly and slowly and are functionally related to one another. A move in one direction of a part may entail the movement of another part in another direction. Hence, the psychiatric treatment itself is a process that the clinician must embark on together with the person, and that forces the clinician to think about system dynamics and the way these contextualize the outcomes of trial-and-error interventions characteristic of practice in psychiatry. One way to appreciate the value of the model presented in chapter 4 is as an additional tool that psychiatrists may use to inform or complement their decisions during this trial-and-error process. Accordingly, the goal of the sort of model proposed in chapter 4 was not to provide a “literally true" map of depression, which would probably be impossible. Rather, the goal was to start developing a tool that could one day accompany clinicians in the healing process. For instance, such a model could support the clinicians in their prioritization of interventions within level and across level dynamics, based on their understanding of which of the multiscale parameters may appear to contribute most to the disorder at hand.

Thesis conclusion

In this dissertation, I sought to integrate three distinct ways of thinking about mental disorders that belong to three distinct approaches to psychiatry: the evolutionary approach, the cultural approach and the computational approach. To do so, I proposed an integrative model of the three approaches in chapter 3, which was based on some assumptions that were derived from the discussions in chapters 1 and 2. The first assumption was that it would be valid to extend to non-living entities the principle at the core of the computational approach underwriting the model of chapter 3 — the free-energy principle (chapter 1). The challenge was to extend the free-energy principle to non-living entities without having to commit to the idea that such non-living entities are in fact living. To justify this, I argued that the free-energy principle was not a sufficient condition for life and cognition; hence one could perfectly well apply the free-energy principle to non-living entities without committing to the view that these entities are living, in some sense. The lesson is that if one is to inquire about what life-related processes are, it may be worth doing this under the auspices of the free-energy principle. However, it is not because one uses the free-energy principle that one is inquiring about life-related processes.

The second assumption was that one could extend applications of the free-energy principle to non-living entities by attributing to both living and non-living entities the same general ability to 'learn' from each other, 'perceive' one another, and 'act' to shape one another's 'perception'. The defence of this assumption was central to our project in chapter 3, which was in part to extend the free-energy principle to model the behaviour of medical clinical interactions and institutions. The view of learning, perception and action as inference processes in Chapter 2 yielded a symmetrical and asymmetrical view of human-environment interactions that would provide a mechanism for the manner in which scales in multilevel models of human behaviour can bind together and become dependent on one another, while allowing for independent dynamics at each scale. I suggested that asymmetrical interactions between scales yield unstable patterns at the level below (e.g., an individual in relation to her environment), whereas the accumulation of such asymmetrical interactions by multiple individuals "alike" at the level above would yield stable patterns at the level above (e.g., the stability of intergenerational passing of tradition that guarantees survival or reproductive success, and upon which individuals depend). The concepts of symmetry and asymmetry in a multiscale system and the stable and unstable patterns they suppose can guide us when attempting to explain the behaviour we observe at a given scale. One should expect symmetry and stability in a scale-to-scale relation (e.g., one should expect fidelity in the reproduction of certain alleles in a population over generations), and asymmetry and instability in the relation between an entity within its own scale and the scale above (e.g., germline mutations from parents to offspring creating variations in the same population).

Finally, in chapter 4, we presented a model of depressed mood and normative depression by implementing the model presented in chapter 3. It is important to note that in this dissertation, even though I have been talking about one model, there are two models at play. The first is the one developed in chapter 3 — the ECC model, which is, as we proposed in introduction, something like a conceptual model for psychiatry — a medical model, like the biopsychosocial model. Such a model can play various roles. It can orient education by providing a set of core principles that can be taught to undergraduate students, e.g., as with evolutionary medicine (Grunspan et al., 2019); it can orient research by integrating various bodies of knowledge as we did in chapter 3; and it can shape clinical practice by helping the clinician to "consider information from all systems' levels and the possible relevance and usefulness of data from each level for the patient's further study and care" (Engel, 1981, p. 101). The second model discussed in this dissertation is that of chapter 4. That model can be viewed as the sort of research output that we can expect under the conceptual model developed in chapter 3.

Thesis references

Andrews, Mel. 2021. "The Math Is Not the Territory: Navigating the Free Energy Principle." *Biology & Philosophy* 36 (3): 30. https://doi.org/10.1007/s10539-021-09807-0.

Badcock, P. B., Davey, C. G., Whittle, S., Allen, N. B., & Friston, K. J. (2017). The Depressed Brain: An Evolutionary Systems Theory. *Trends in Cognitive Sciences*, *21*(3), 182–194.

Boorse, C. (1977). Health as a Theoretical Concept. *Philosophy of Science*, *44*(4), 542–573.

Burcusa, S. L., & Iacono, W. G. (2007). Risk for recurrence in depression. *Clinical Psychology Review*, *27*(8), 959–985.

Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis*, *58*(1), 7–19.
Engel, G. L. (1981). The Clinical Application of the Biopsychosocial Model. In *Journal of Medicine and Philosophy* (Vol. 6, Issue 2, pp. 101–124). https://doi.org/10.1093/jmp/6.2.101
Fleck, L. (1979). *Genesis and Development of a Scientific Fact*. University of Chicago Press.
Friston, K. (2017). I am therefore I think. In M. Leuzinger-Bohleber, S. Arnold & M. Solms (Eds.), The unconscious : a bridge between psychoanalysis and cognitive neuroscience (pp. 113-137). London: Routledge.
Friston, K., Levin, M., Sengupta, B., & Pezzulo, G. (2015). Knowing one's place: a free-energy approach to pattern regulation. *Journal of the Royal Society, Interface*, *12*(105), 20141383. https://doi.org/10.1098/rsif.2014.1383
Giroux, É. (2015). Epidemiology and the bio-statistical theory of disease: a challenging perspective. *Theoretical Medicine and Bioethics*, *36*(3), 175–195.
Godfrey-Smith, P. (1994). A Modern History Theory of Functions. *Noûs*, *28*(3), 344–362.
Griffiths, P. E. (1993). Functional Analysis and Proper Functions. *The British Journal for the Philosophy of Science*, *44*(3), 409–422.
Griffiths, P. E., & Matthewson, J. (2018). Evolution, Dysfunction, and Disease: A Reappraisal. In *The British Journal for the Philosophy of Science* (Vol. 69, Issue 2, pp. 301–327). https://doi.org/10.1093/bjps/axw021
Grunspan, D. Z., Moeller, K. T., Nesse, R. M., & Brownell, S. E. (2019). The state of evolutionary medicine in undergraduate education. *Evolution, Medicine, and Public Health*, *2019*(1), 82–92.
Huda, A. (2019). *The Medical Model in Mental Health: An Explanation and Evaluation*. Oxford University Press.
Kirmayer, L. J., & Young, A. (1999). Culture and context in the evolutionary concept of mental disorder. *Journal of Abnormal Psychology*, *108*(3), 446–452.
Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.
Latour, B. (2000). *Pandora's hope : essays on the reality of science studies*. Harvard University Press.
Matthewson, J., & Griffiths, P. E. (2017). Biological Criteria of Disease: Four Ways of Going Wrong. *The Journal of Medicine and Philosophy*, *42*(4), 447–466.
Millikan, R. G. (1984). *Language, Thought and Other Biological Categories*. MIT Press.
Morris, Sarah E., Charles A. Sanislow, Jenni Pacheco, Uma Vaidyanathan, Joshua A. Gordon, and Bruce N. Cuthbert. 2022. "Revisiting the Seven Pillars of RDoC." *BMC Medicine* 20 (1): 220. https://doi.org/10.1186/s12916-022-02414-0.
Nesse, R. M. (2019). *Good Reasons for Bad Feelings: Insights from the Frontier of Evolutionary Psychiatry*. Dutton.
Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active Inference: The free-energy Principle in Mind, Brain, and Behaviour*. MIT Press.
Pickering, A. (1995). *The mangle of practice: time, agency, and science*. University of Chicago Press.
Schwartz, P. H. (2007). Defining Dysfunction: Natural Selection, Design, and Drawing a Line. *Philosophy of Science*, *74*(3), 364–385.
Siegle, Greg J., Angélique O. J. Cramer, Nees Jan van Eck, Philip Spinhoven, Steven D. Hollon, Johan Ormel, Marlene Strege, and Claudi L. H. Bockting. 2019. "Where Are the Breaks in Translation from Theory to Clinical Practice (and Back) in Addressing Depression? An Empirical Graph-Theoretic Approach." *Psychological Medicine* 49 (16): 2681–91. https://doi.org/10.1017/S003329171800363X.
Sims, Matthew. 2021. "How to Count Biological Minds: Symbiosis, the Free Energy Principle, and Reciprocal Multiscale Integration." *Synthese* 199 (1): 2157–79. https://doi.org/10.1007/s11229-020-02876-w.
Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World*. OUP USA.
Whiten, A., & Erdal, D. (2012). The human socio-cognitive niche and its evolutionary origins. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1599), 2119–2129.

[i] The Evolutionary System Theory (EST) reading of depression as resulting from an aggravated vulnerability having provided selective advantage is one possible evolutionary reading. As Nesse (2019) put it: "Every trait or gene that makes an organism vulnerable to disease poses an evolutionary mystery. The old answer was that there are limits to what natural selection can do—for instance, eliminating all mutations. That is one important kind of explanation, but the central insight of evolutionary medicine is that there are

also at least five other evolutionary reasons why we are vulnerable to diseases" (Nesse, 2019, p.34). The reasons behind the maintenance of vulnerabilities underwriting major depression are most probably more complex than the simple fact that "they might have provided a selective advantage". These reasons, for instance, may, and in fact, should include, as Nesse suggests, the evolutionary rationales described in chapter 3. According to Nesse (2019): "Genes or traits associated with some diseases provide advantages and disadvantages that influence natural selection. However, proposals about the utility of diseases themselves, such as schizophrenia, addiction, autism, and bipolar disorders, are wrong before they start. The correct question is Why did natural selection shape traits that make us vulnerable to disease? Such vulnerabilities need an evolutionary explanation using some combination of these six factors." (Nesse, 2019, p.41); the "six factors" being the evolutionary rationales summarized by Nesse (2019), some of which overlap with the ones described in chapter 3 of this dissertation. In chapter 4, the simpler, naïve account is used because it allows us to implement the EST of depression in a simple computational model. Within the context of this dissertation, chapter 4 functions as a "proof of principle" for the evolutionary, cultural, and computational approach developed in chapter 3. That proof of principle is about whether one could develop a sound computational model of psychiatric phenotypes reflecting dynamics that conform to principles of evolutionary, cultural, and computational psychiatry. The proposed model in chapter 4 is most certainly miles away from the "literal truth". But this should not be surprising. The work in chapter 4 is very much exploratory.

[ii] See end note 1.