



City Research Online

City, University of London Institutional Repository

Citation: Rigoli, F. (2022). Prisoner of the present: Borderline personality and a tendency to overweight cues during Bayesian inference. *Personality Disorders: Theory, Research, and Treatment*, 13(6), pp. 609-618. doi: 10.1037/per0000549

This is the accepted version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: <https://openaccess.city.ac.uk/id/eprint/27597/>

Link to published version: <https://doi.org/10.1037/per0000549>

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

City Research Online:

<http://openaccess.city.ac.uk/>

publications@city.ac.uk

**Prisoner of the present: Borderline personality and a tendency to overweight cues
during Bayesian inference**

Francesco Rigoli, City University

Abstract

Recent work has examined the computational mechanisms underlying Borderline Personality (BP). However, this research has been confined to specific tasks. A computational analysis of BP's mental processes as they broadly unfold in everyday life is lacking. Here a computational model of BP is proposed which describes patients' everyday-life mental experience at large. Grounded on Bayesian inference, the proposal is that BP sufferers attribute excessive weight to cues considered to infer life contexts (e.g., to infer whether a cooperation or competition context is ongoing). Remarkably, model simulations demonstrate that this idea accounts for several characteristics of BP, from extreme oscillations in identity, affect, and behaviour, to dysfunctional interpersonal cycles. Altogether, the paper offers a framework to interpret the broad, everyday life computational mechanisms underlying BP. This can inspire theoretical and empirical research and can help understanding how clinical interventions for BP work, thus contributing to refine such interventions.

Keywords: Borderline personality; computational psychiatry; Bayesian; inference; life context; mutual information

Introduction

Borderline personality (BP) is a serious mental disorder affecting from 1.2% to 6% of individuals (Grant et al., 2008; Lieb et al., 2004). Besides the severe psychological distress, its costs comprise the burden for health services (in the USA, more than 20% of people accessing clinical inpatient services hold this diagnosis) and, most dramatically, the extremely high rate of suicide among its sufferers (Grant et al., 2008; Lieb et al., 2004). Being a disorder of personality, BP is general and stable, as it encompasses several life domains and persists over time, often resisting clinical intervention. Despite a remarkable heterogeneity (according to standard diagnosis criteria, as many as nine symptoms can be manifested, with only five of them being enough for a diagnosis (American Psychiatric Association, 2013)), all BP sufferers share the same abrupt fluctuations in affect, beliefs, and identity, so much so that rapidly alternating extremes (in affect, beliefs, and identity) are considered to be the core of the illness (Lieb et al., 2004; Linehan, 1993). For example, in a remarkably short time interval, the same BP patient might go from expressing excessive euphoria to manifesting extreme anger or sadness, usually to the perplexity of others. Peaks in anger, sadness or fear are particularly problematic because they often result in impulsive actions such as in violence towards others or towards oneself (in the latter case, not rarely manifested in suicidal behaviour). Moreover, chronic psychological instability renders intimate relationships intense but highly unstable, resulting frequently in breaking points.

What are the mental processes responsible for the development and maintenance of BP? Several lines of research have offered valuable insight. Among these, computational psychiatry is a recent approach which interprets mental illness in terms of computational mechanisms gone awry (Huys et al., 2016; Montague et al., 2012). By relying on formal mathematical modelling, this approach offers a precise description of the processes implicated in the formation and maintenance of a disorder, providing clear definitions of classical concepts adopted in the clinical literature and of their relation. Though this approach comes with the price of making substantial simplifications, many have argued that this is more than compensated by the clarity afforded by mathematics, which facilitates both theoretical debate and identification of specific empirical predictions (Huys et al., 2016; Montague et al., 2012).

Relying on a computational psychiatry perspective, recent work has asked BP patients to perform judgement and decision-making tasks which have been analysed adopting computational modelling (Fineberg, et al., 2017; 2018; Franzen et al., 2011; Henco et al., 2020; King-Casas et al., 2008; Siegel et al., 2020; Unoka et al., 2009). This research has pinpointed to specific aspects distinguishing BP sufferers from non-sufferers, so that a picture is emerging about which computational mechanisms might underly BP. However, so far, the scope of this research has been somewhat confined to the context of specific tasks. An analysis of the broader implications for understanding mental processes of BP as they unfold in everyday life remains to be developed. In other words, what can we learn from recent computational psychiatry research in BP in terms of how this disorder affects everyday experience? Can we build upon this research to develop a general picture of BP in terms of how patients behave in their ecological contexts, of how they shape their relationships, and of how they make choices in real life? Such broad, ecological theories of BP exist (e.g., Fonagy et al., 2000; Kernberg, 1967; Linehan, 1993; Liotti, 2002), but they are not computational. In an attempt to bridge ecological theories of BP with recent computational psychiatry research on this illness, here I propose an ecological computational model of BP. The purpose of this is to adopt computational modelling to develop a theory of BP which encompasses patients' everyday life experience at large.

The next section overviews the theory, which is referred to as Ecological Computational Model of Borderline Personality (ECMBP) (here the word *ecological* refers to an attempt to describe patients in their every-day life contexts, beyond any specific psychological task). After outlining the theory, this is analysed in a set of model simulations exploring how typical clinical manifestations of BP arise when specific model parameters are set. Finally, the model is discussed in the context of broader issues.

The Model

The ECMBP is based on Bayesian statistics (Bishop, 2006). The assumption is that the brain represents variables that are key for interpreting everyday life experience and that, based on observing

some of them, it infers the state of others (Oaksford & Chater, 2007). The model can be described adopting a Bayesian network formalism (Bishop, 2006) where circles and arrows represent probabilistic variables (all categorical in the ECMBP) and their relationships, respectively (fig. 1). At the centre of this network is the variable Current Life Context (LC_C). The proposal is that the brain parcels everyday life experience in discrete categories or life contexts. A life context is defined by specific manifestations for each of three main dimensions including: (i) self-identity, capturing beliefs about expectations, affects, goals, and behaviours of the self in a given context, (ii) others-identity, capturing beliefs about expectations, affects, goals, and behaviours of others in a given context (with one or more individuals being at play) (note that some life contexts might not involve interactions with others; for these contexts, the others-identity dimension is not relevant), and (iii) environment-identity, capturing beliefs about the physical setting where an action or interaction takes place. Analogous to previous concepts in the literature such as those of internal working model and self-schema (Bretherthon & Munhollad, 2008; Kendzierski, 1980; Markus, 1977), the notion of life context describes an abstract representation which pinpoints to general (personal, interpersonal, or social) structures and dynamics that can be experienced in everyday life. We remain agnostic about how many and which life contexts might be available to the brain as options, not least because this is likely to depend strongly on each individual life history. At the same time, it is arguable that some life contexts might be a recurrent option among people. For example, an attachment context, where the self is interpreted as in need of protection and the other as a caring figure, might be an option universally available (Bretherthon & Munhollad, 2008). Another example might be a competition context, viewing the self and the other as similar in status and in competition with one another.

The ECMBP proposes that, during development (and possibly under the influence of genetic predispositions) the brain learns to categorise the multifarious manifestations of everyday experience based on a relatively short list of life contexts. In other words, the brain would assume that, although everyday experience is often diverse and confused, ultimately this can be filtered out by invoking one among a small set of life contexts which would be at play in the present. The ongoing life context (the LC_C), the proposal goes, would not be observable directly (being a hidden or latent variable), but can

be inferred indirectly based on related cues (a cue can be any stimulus, or complex of stimuli, in the environment). For example, facial expressions manifested by another person (e.g., an aggressive face) might be interpreted as a cue reflecting a specific underlying life context (e.g., a competition context). Although the ECMBP can include multiple cue variables, for the sake of simplicity only three are implemented in the graphical model of fig. 1 (C_A , C_B , and C_M). Whereas C_A and C_B reflect cues from the external physical or social environment, C_M captures information about the own behavioural or psychological processes. For instance, it might indicate which episodic memory trace is currently retrieved (e.g., the idea being that remembering a fight with a classmate would signal that a competition life context is ongoing).

Finally, the graphical model described in fig. 1 includes a variable reflecting the Past Life Context (LC_P), capturing the life context at play in the recent past. The idea is that the brain assumes that the past influences the present in such a way that, if a certain life context was active in the immediate past, there is a good chance that it remains active now.

Formally, the joint probability of the variables included in the graphical model corresponds to:

$$P(LC_P, LC_C, C_A, C_B, C_M) = P(LC_P) P(LC_C | LC_P) P(C_A | LC_C) P(C_B | LC_C) P(C_M | LC_C)$$

This implicates the following generative process. First, a past life context is sampled from LC_P with probability $P(LC_P)$. Second, the current life context is sampled from the conditional probability $P(LC_C | LC_P)$. Finally, cues signalling the current life context are sampled from $P(C_A | LC_C)$, $P(C_B | LC_C)$, and $P(C_M | LC_C)$. Any cue variable can be observed and be relied upon to perform Bayesian inference. Inference can concern the posterior probability of the current life context (e.g., if both C_A and C_B are observed, these can be considered to infer $P(LC_C | C_A, C_B)$) or can concern the posterior probability of a cue variable which has not been observed (e.g., if C_A alone is observed, this can be considered to infer $P(C_B | C_A)$). Notably, the Bayesian graph in fig. 1 can be interpreted as a Markovian process, and thus adopted to make inference over different time points (formally, resulting in a Hidden Markov Model; this is analogous to a Kalman filter, although the latter applies to continuous variables while the ECMPB focuses on categorical variables; Bishop, 2006). This occurs simply by treating LC_C at

time t as equal to LC_P at time $t+1$: once the posterior of LC_C (e.g., $P(LC_C | C_A, C_B)$) is calculated for time t , this can be considered to be equal to $P(LC_P)$ at time $t+1$, which in turn can be used to infer the posterior of the new LC_C at time $t+1$.

What we have overviewed so far unites BP sufferers and non-sufferers alike. So, what does distinguish the two groups? We propose that the key difference relies on the weight attributed to cues. Cues can be viewed as more or less informative during inference. For some people, cue variables might be considered as highly informative, namely their knowledge might really make a difference during inference. For other people, knowing cue variables might make only marginal difference during inference. Formally, the weight of any cue variable C_X can be quantified by the Mutual Information MI between C_X and LC_C (Dayan & Abbott, 2001):

$$MI(C_X, LC_C) = \sum_{LC_C} \sum_{C_X} P(LC_C) P(C_X | LC_C) \log \frac{P(C_X | LC_C)}{P(C_X)}$$

The Mutual Information increases when knowing C_X is more informative in terms of inferring LC_C . The ECMBP proposes that the key difference between BP sufferers and non-sufferers is that the former weight cue variables more; in other words, the Mutual Information between any cue variable C_X and LC_C is proposed to be higher in BP. This idea offers an interpretation for a variety of empirical data. First, it is consistent with the observation that, when making judgements in inference and decision-making tasks, BP patients rely more than controls on ongoing cues (Fineberg et al., 2018). Second, research has revealed a strong link between situational triggers and symptoms in BP (Miskewicz et al., 2015), an observation that can be interpreted as arising from relying heavily on situational cues. Third, BP is linked with enhanced absorption (i.e., a tendency to immerse oneself within ongoing stimulation) (Zanarini et al., 2000; Zweig-Frank 1994a; 1994b), a phenomenon that can be formalised in terms of heightened sensitivity to ongoing cues. Fourth, recent evidence suggests that BP patients are more susceptible to perceptual illusions which depend on attributing higher emphasis on visual cues (Neustadter et al., 2019).

The notion of excessive cue-weighting also presents analogies with recent theories of BP, framing these within a computational outlook. An influential perspective adopts network theory to explain mental disorders (including BP), interpreted as characterised by strong (vs. weak) network connectivity (Borsboom, 2017; Burger et al., 2020). The idea of large mutual information between cues and contexts in BP can be interpreted as an instance of strong connectivity between two elements of the network. Postulating an exaggerated susceptibility to triggers as being at the core of BP, The Symptom-Trigger Contingency model is another recent influential perspective on the disorder (Miskewicz et al., 2015). The ECMBP offers a computational interpretation of the notion of trigger susceptibility, casted in terms of cue-weighting during inference. A third recent proposal explains BP as arising from an enhanced influence of Pavlovian cues during Pavlovian-to-instrumental transfer (Hallquist et al., 2018). The ECMBP provides a computational formulation of this idea, and suggests that an enhanced cue-influence might be more general (not restricted to Pavlovian-to-instrumental transfer). Below, broad implications of attributing higher weight to cues, here proposed to be at the core of BP, are explored through model simulations.

Simulations

Simulation 1: inferring life contexts

Consider a scenario where two alternative life contexts are available as options: a Cooperation context (where the self and the other collaborate honestly to achieve a common goal) and a Betrayal context (where the other betrays the self in order to achieve goals at the expense of the self's goals). The two contexts represent the categories available for both LC_C and LC_P . One single cue variable (C_A) is available in this scenario, with categories being Betrayal cue (e.g., indicating that the other person is keeping some resources for herself) versus Cooperation cue (e.g., indicating that the other person is sharing resources). The former category is believed to be more likely given the Betrayal context, and the latter category given the Cooperation context. This scenario is simulated for several trials, being a new cue C_A sampled at every trial. Let us assume that, throughout the whole temporal sequence, the true intention of the other person is to cooperate, meaning that the Cooperation context is the actual

life context at play all along. However (being observations uncertain by nature), although Cooperation cues are usually gathered, sporadically Betrayal cues appear (e.g., the other person might occasionally perform seemingly betraying actions because of a temporary lapse, not because of a real intention of betraying). Adopting this scenario, let us compare inference among three individuals characterised by varying degrees of Mutual Information between LC_C and C_A . In accordance with the ECMBP, the person with the highest Mutual Information is interpreted as being affected by BP. The top panel of Fig. 2 plots the posterior $P(LC_C | C_A)$ over trials as inferred by each individual. When a Betrayal cue appears, control individuals still judge the Cooperation context as the most likely; correctly, they interpret the Betrayal cue as due to noise, and give it low importance. Conversely, when occasionally presented with a Betrayal cue, a Betrayal context is inferred as the most likely by the BP agent. The Betrayal cue is attributed an unwarranted importance, as it is not interpreted as noise but as reflecting a real change in life context. This example shows a key impairment derived from attributing an excessive weight to ongoing cues: noisy observations are not filtered out and thus are taken too seriously. The implication is that, although the life context in fact does not change, BP individuals' beliefs about the ongoing life context oscillate dramatically. This captures one of the key aspects of BP, namely the frequent and often puzzling swinging of beliefs, mood, and identity (Lieb et al., 2004; Linehan, 1993). Moreover, this scenario fits with empirical evidence showing that, when partaking in cooperation games, BP patients tend to break cooperation following misunderstandings with confederates (King-Casas et al., 2008). The ECMBP interprets this as arising because BP patients attribute excessive weight to cues, so that occasional non-cooperation signals are given excessive importance. It has also been found that, in judgement and decision-making tasks, BP patients manifest lower learning rate, namely they appear to be less influenced by the past (Fineberg et al., 2018; Henco et al., 2020). This is consistent with the scenario examined here, where inference expressed by the BP agent is driven primarily by ongoing cues and not by past beliefs (captured by LC_P).

The bottom panel of fig. 2 plots the Entropy (H) of the posterior probability $P(LC_C | C_A)$ along trials, equal to (Dayan & Abbott, 2001):

$$H(LC_C | C_A) = - \sum_{LC_C} P(LC_C | C_A) \log(LC_C | C_A)$$

The entropy (which is equal to zero when one category has probability of one and other categories have probability of zero; and which is maximal when all categories have equal probability) reflects the uncertainty about the posterior life context. Fig. 2B indicates that the entropy is generally smaller for the BP agent compared to controls. In other words, the BP agent is predicted to be more confident (i.e., less uncertain) about inference concerning life contexts. In general, because of an excessive weight attributed to cues, the ECMBP predicts that BP sufferers will be more confident about their inferences concerning life contexts. This prediction fits with lab observations indicating that BP patients report higher confidence in their social judgements (Shilling et al., 2012; Siegel et al. 2020).

Altogether, both fig. 2A and 2B indicate that, as a consequence of weighting cues heavily, the BP agent manifests extreme inferences (i.e., the life context inferred as the most likely is attributed higher posterior probability). What does this imply for behaviour? Let us assume that (i) a person performs actions according to the life context inferred as the most likely (e.g., a person acts cooperatively if the Cooperation life context is judged as the most likely, and she acts defensively if the Betrayal life context is judged as the most likely) and that (ii) the vigour of actions increases as entropy decreases (i.e., when confidence about the life context increases) (Rigoli, 2021). Being linked with lower entropy, extreme inferences as those expressed by the BP agent imply higher action vigour. This explains empirical evidence showing that BP patients exhibit extreme emotional responses (Linehan, 1993): higher confidence about life contexts would imply more vigorous emotional reactions (e.g., higher certainty of being in a Betrayal context would elicit stronger anger; note that, as empirical evidence suggests (Lerner & Keltner, 2001), such mechanism might be particularly salient when experiencing anger). Furthermore, this explains evidence of greater behavioural impulsivity in BP (Crowell et al., 2009; Lieb et al., 2004; Linehan, 1993): higher confidence about life contexts would imply less cautious behaviour and greater propensity to act out (e.g., higher confidence of being in a Betrayal context would motivate ready and impulsive retaliation behaviour).

The scenario examined here can shed light on yet another aspect of the illness. There is abundant evidence of impaired theory of mind among BP sufferers (Fonagy & Bateman, 2007; Nemeth et al., 2018; Semerari et al., 2005; Sharp et al., 2011). These people often report distorted understanding of the psychological processes expressed by themselves and by others. The ECMBP offers a formalization of theory of mind deficits in BP, which are explained as distorted beliefs about how mental processes are displayed. More specifically, the notion of life context encompasses beliefs about mental processes (i.e., expectations, affect, and identity) guiding oneself and others. The conditional probability $P(C_A | LC_C)$ reflects beliefs about how such mental processes are displayed through cues (e.g., through emotional or behavioural signals). The ECMBP proposes that, in BP, $P(C_A | LC_C)$ is distorted because cues are attributed excessive importance. For example, BP patients might wrongly interpret someone not paying attention as revealing an underlying aggressive intension, or someone being polite as revealing an underlying erotic attraction. Such inability to interpret cues correctly and attribute them the appropriate weight is, according to the ECMBP, a main theory of mind deficit in BP. As examined above, a consequence of this is an excessive confidence about judgements. This can be interpreted as a second major theory of mind deficit proposed by the ECMBP to describe BP: a well-functioning theory of mind requires an accurate level of confidence about one's own judgements, while BP patients would manifest excessive confidence (Shilling et al., 2012; Siegel et al., 2020).

In summary, this simulation elucidates key implications of the ECMBP, a model proposing that BP patients weight cues excessively. This implies that patients give too much importance to cues due to noise, thus manifesting dramatic oscillations in their beliefs, more extreme beliefs, and enhanced confidence in their inferences. The explanatory power of just this single aspect is remarkable, as it can account for various features of BP as observed empirically.

Simulation 2: inferring cues

This simulation is like the first one except that now a second cue variable (C_B) is included in the model. For example, C_B might describe the behaviour of a third person, with categories being a Cooperation cue (e.g., when the person displays impartial cooperation) versus Betrayal cue (when the third person favours the second person), the former and the latter supporting the Cooperation and Betrayal life context, respectively. However, despite being included, C_B is never observed; rather, at every trial this cue variable is now inferred based on observing C_A and on calculating the posterior probability $P(C_B | C_A)$. Intuitively, this inference represents a guess about an event which in principle could have been recorded (e.g., the behaviour of the third person could have been observed), though in fact it was not. The top panel of fig. 3 describes the posterior $P(C_B | C_A)$ along trials, again for three individuals characterised by varying degrees of Mutual Information between LC_C and C_A (which is assumed to be equal to the Mutual Information between LC_C and C_B). As above, the individual with the highest Mutual Information is labelled as BP. Fig. 3 illustrates that $P(C_B | C_A)$ oscillates dramatically for the BP agent, insofar as observing sporadic Betrayal cues for C_A leads to inferring Betrayal cues also for C_B (e.g., it leads to guessing that the third person is favouring the second person). This does not occur in control agents, where a Cooperation cue for C_B continues to be guessed even following presentation of occasional Betrayal cues for C_A . The bottom panel of fig. 3B displays the Entropy (H) of $P(C_B | C_A)$ along trials, which is equal to:

$$H(C_B | C_A) = - \sum_{C_B} P(C_B | C_A) \log(C_B | C_A)$$

As Fig. 3 indicates, the BP agent displays lower entropy (i.e., higher confidence) concerning its inference about C_B (e.g., she is more confident about the third person's behaviour).

This scenario interprets two phenomena characterising BP. First, it fits with evidence showing that BP sufferers sometimes experience temporary delusions, as they manifest absolute conviction about the occurrence of an event (e.g., they might be certain of being cheated or of being chased) which is in fact unlikely (D'Agostino et al., 2019; Pearse et al., 2014). The ECMBP explains temporary delusions

in BP as the effect of weighting cues excessively during inferences of a cue variable from other cue variables.

The second phenomenon captured by this scenario concerns internal cues. As examined above, while some variables (e.g., C_A and C_B) reflect external cues, other variables (e.g., C_M) reflect signals coming from the own body or mind. For instance, C_M might describe memory traces (e.g., the recollection of playing with a friend, or of being deceived by another). Sometimes, C_M might be observed (e.g., a memory trace might be currently active) and drive inference of life contexts like any other cue variable (the logic being that, if one is remembering a certain event, then the associated life context is more likely). Other times, C_M might be the target of an inference based on other cues, like in the scenario examined here. The outcome of this might be the retrieval of a specific memory trace: the posterior probability $P(C_M | C_A)$ might be inferred and one category of C_M might be retrieved from memory with a chance equal to its posterior probability. Although this favours retrieval of the trace associated with the highest posterior probability, for control individuals other traces still have a substantial chance of being retrieved. Conversely, for BP, the ECMBP implies that one memory trace will have excessive posterior probability, thus blocking retrieval of other traces. Consequently, in BP memory retrieval is predicted to be monopolised by the trace associated with the highest posterior probability. This explains empirical data about temporary memory deficits in BP, as patients often struggle to remember events which are inconsistent with their current thoughts or mood (Jones 1999; Winter et al., 2014).

In summary, this simulation explains how, according to the ECMBP, events are conjectured by BP patients. This is captured by an inference where a cue variable is estimated from another cue variable. This form of inference might underly BP's characteristics such as the occurrence of transitory delusions and congruency effects in memory retrieval.

Simulation 3: interpersonal cycles

The interpersonal sphere is critical for understanding BP (Dimaggio et al., 2007; Fonagy & Luyten, 2009; Liotti, 2014; Semerari & Fiore, 2007). Patients suffering from this disorder experience problems both in intimate and occasional social interactions, due to the extreme and abrupt oscillations in emotion and behaviour they manifest (Sadikaj et al., 2013). Thus, given the centrality of interpersonal dynamics in BP, this simulation assesses the ECMBP in this domain.

The simulation considers pairs of interacting agents, where each agent is equipped with a Bayesian network including LC_P , LC_C , and C_A . LC_P and LC_C each include three categories: Cooperation, Competition, and Neglect (the latter being a context where none of the agents is caring about the other). C_A describes the action performed by the other agent and, for each life context, includes two forms of action (six in total), one associated with high vigour and the other associated with mild vigour. Thus, C_A indicates whether the other agent has performed a vigorous or mild Cooperation action, a vigorous or mild Competition action, a vigorous or mild Neglect action. $P(C_A | LC_C)$ describes beliefs about the probability of observing a certain action performed by the other agent given the ongoing life context. For example, if a Cooperation context is ongoing, then the other agent is predicted to be more likely to exhibit a (vigorous or mild) Cooperation action. On every trial t , each agent performs one of the six available actions. On the following trial $t+1$, this action is recorded by the C_A of the other agent who, on this basis, infers the posterior life context $P(LC_C | C_A)$ for $t+1$. Once this has been inferred, an agent decides which action to perform at $t+1$. This decision works as follows. First, the inferred life context (i.e., the one with the highest posterior probability for $P(LC_C | C_A)$ at time t) can be selected with a fixed probability (in the simulation, equal to 0.8). If this context is selected, then the corresponding action will be performed (e.g., if the cooperation context is selected, either a vigorous or mild cooperation action will be performed; specifically, a vigorous action is performed when the entropy of $P(LC_C | C_A)$ is lower than a threshold – in the simulation equal to 0.6). If the inferred life context is not selected, then any mild action associated with any of the three life contexts can be performed, each action with equal probability. When an action is

performed at trial $t+1$ by an agent, at trial $t+2$ the action is recorded by the C_A of the other agent who uses it to infer $P(LC_C | C_A)$ for $t+2$, and the cycle is repeated.

We simulate two pairs of agents interacting over several trials. The first pair includes two control agents, both characterised by low Mutual Information between LC_C and C_A . The second pair includes one control agent interacting with a BP agent, the latter characterised by high Mutual Information between LC_C and C_A . Both pairs start in a Cooperation life context (this is obtained by assigning higher probability to the Cooperation LC_P at time 1).

The left panels of fig. 4 describe the inferred life context (i.e., the context judged as the most likely a posteriori) for each agent of the first pair. For both agents, a Cooperation context persists for as long as 70 trials, when it is replaced by a Competition context lasting until the end. Overall, contexts appear rather stable: only two contexts are experienced by the pair. Moreover, when a specific context is activated, both agents appear stable in their inferences, as they usually interpret occasional inconsistent behaviour as noise, and not as an actual shift in context.

The right panels of fig. 4 describe the inferred life context for each agent of the second pair (including a BP agent). After around 40 trials, both agents shift from a Cooperation to a Neglect context. The latter lasts only for a few trials when it is replaced by a Competition context. Altogether, as many as 8 contexts alternate, highlighting a marked instability. Moreover, consider the BP agent during trial 50 to 90. Within this interval, although a Competition context is usually inferred by the BP agent, inference appears as markedly unstable, in as much as several trials are attributed to the Neglect context. Overall, interactions appear as highly unstable for the second pair of agents. Why? This derives from the BP's tendency to interpret occasional behavioural inconsistencies not as noise, but as real context shifts. When the BP agent infers a context shift, she reacts accordingly, and this reaction is in turn recorded by the other agent. Insofar as now both agents believe that the context has changed and act accordingly, the final result will be an actual context shift. This is a form of self-fulfilling prophecy: it is the BP agent's conviction that the context has changed (leading the BP agent to act accordingly and eliciting a reaction by the other agent) that ultimately causes an actual change in

context. The new context lasts until, again, by chance the other agent performs an inconsistent behaviour which is interpreted by the BP agent as a true context shift. This simulation offers a computational analysis of the nature of dysfunctional interpersonal cycles experienced by BP patients. The ECMBP proposes that, because of an excessive weight attributed to cues, BP patients often misinterpret others' actions, leading to frequent shifts in how they interact with others. These dramatic interpersonal shifts would prevent the dyad to maintain stable interactions. Not surprisingly, despite the best of effort, experiencing such unstable interpersonal cycles would lead many people not affected by BP to break their relationship with BP patients (Dimaggio et al., 2007; Fonagy & Luyten, 2009; Liotti, 2014; Semerari & Fiore, 2007).

In summary, this simulation examines the computational mechanisms at the root of interpersonal dysfunctions characterising BP. The picture is consistent with a view of BP as characterised by an ambiguous (within-situation) interpersonal pattern which, depending on how the other person acts, can result in very different behaviours, producing a sort of *reliable instability* (Schmideberg, 1959; Hopwood, 2018). Notably, the same simple idea proposed above (that overweighting cues is at the core of BP) offers insight also on the nature of interpersonal dysfunctions in the disorder.

Discussion

The paper introduces the ECMBP, a computational theory of BP which attempts to describe the disorder globally and ecologically, going beyond the processes at play in specific tasks or circumstances. The key idea is simple: at the heart of BP is proposed to be an excessive emphasis on cues. Remarkably, this idea can account for a variety of characteristics of BP, from the manifestation of extreme oscillations in identity, affect, and behaviour, to dysfunctional interpersonal cycles.

This proposal raises an obvious question: where would the alleged overemphasis on cues come from in BP? Though a systematic answer goes beyond the scope of the manuscript, two broad influences can be postulated: genetic factors (some individuals might be genetically predisposed to rely heavily on available cues) and past experience. Bayesian psychological theories often presuppose that, besides

genetic influences that might constitute important constraints, a model employed in the present reflects experience collected in the past (Oaksford & Chater, 2007). Psychopathology would ensue when a dysfunctional model is built upon abnormal past experience and is translated into the present, when now experience is not abnormal anymore. Applying this logic to BP, substantial evidence indicates that neglect, (physical and psychological) abuse, and traumas are common in the infancy of BP sufferers (Fossati et al., 1999; Ball & Links, 2009; Widom et al., 2009). This early experience might favour the development of a model characterised by an overemphasis on cues, possibly because such experience, given its extreme and volatile nature, requires to leverage strongly on available cues to come up with a fast assessment of the situation and deal with it effectively.

The variety of BP characteristics that can be captured by the ECMBP is notable. Is the ECMBP sufficient for explaining all facets of BP? The answer is clearly no, if only because the theory is unable to explain the heterogeneity of the disorder (Grant et al., 2008; Lieb et al., 2004). However, the ECMBP might still offer some insight on such heterogeneity. Arguably, people vary widely regarding the generative model proposed here, namely regarding which life contexts are available as options and regarding their link with cues. This is to be expected also among BP patients; exploring it might offer insight on the computational mechanisms characterising specific subgroups of patients: for example, different subgroups might vary regarding which specific cues are overweighted more than others. Another aspect which is problematic for the ECMBP concerns the discriminant validity of the concept of BP. Recent evidence indicates that, when examined in the context of other personality disorders, BP symptoms do not represent a specific factor, but rather reflect a general dimension (Sharp et al., 2015). Moreover, BP symptoms correlate with all facets of the DSM-5 trait model (Watters et al., 2019), and BP is sometimes indistinguishable from a general factor of psychopathology (Gluschkoff et al., 2021). These observations have led some scholars to suggest that BP should be considered as reflecting the severity of a general dimension of personality disorder, rather than a specific class of personality disorder (Clark et al., 2018). This has implications for the ECMPB: it raises the possibility that the mechanisms proposed by the model might not be limited to a certain class of personality disorders, but rather reflect a general factor characteristic of personality disorders as a whole. This

possibility remains to be examined empirically: the above empirical literature discussed in support of the ECMPB (e.g., Fineberg, et al., 2017; 2018; Franzen et al., 2011; Henco et al., 2020; King-Casas et al., 2008; Siegel et al., 2020; Unoka et al., 2009) does not examine BP in the context of other personality disorders, thus being inappropriate to assess whether the reported effects are specific of BP.

The ECMBP requires nuances also concerning the distinction between positive (e.g., cooperation) and negative (e.g., competition) contexts. As it is now, the ECMBP implies that BP patients will be equally affected by cues belonging to either type of context. However, there is evidence that these patients are more likely to retrieve memories about malevolent individuals (Nigg et al., 1992), and that they fail to update their beliefs about harmful people when positive information is provided (Siegel et al., 2020). This evidence can be reconciled with the ECMBP by postulating an asymmetry between cues belonging to positive and negative contexts, with the latter being weighted more than the former during inference.

Besides explaining available data, the ECMBP aims at inspiring new empirical research. For example, the model proposes that BP's exaggerated cue influence is general, and thus at play across domains; this remains to be assessed empirically. However, the ECMBP does not implicate that all domains are equal: some may reveal higher cue influence than others. In addition, as argued above, different BP patients might vary regarding which contexts are characterised by higher cue influence. In general, domains where BP's cue influence might be particularly high might be those that are particularly salient (e.g., emotionally charged) and those that are less familiar (e.g., novel environments). With this regard, it has been recently shown that people with personality pathology tend to be more confident about the mental states of others than about their own mental states (Müller et al., 2021). Assuming that people are generally less familiar about others than about themselves, the ECMBP can interpret this finding as due to higher cue influence in less familiar domains. Another aspect where the ECMBP can contribute to empirical research concerns the question of how high cue influence develops in the first place. A possibility is that early experience of neglect, (physical and psychological) abuse, and traumas might favour the development of a model characterised by an

overemphasis on cues, possibly because such experience, given its extreme and volatile nature, requires to leverage strongly on available cues to come up with a fast assessment of the situation and deal with it effectively. The link between early trauma and cue influence remains to be explored empirically.

The ECMBP can help understanding how clinical interventions for BP work, potentially contributing to develop better treatments. For example, a well-established rule for treating this disorder is the involvement of multiple therapists for one patient (Bohus et al., 2004). This rule is based on the expectation that, at some point during the therapy, a patient will conflict with each therapist. If one therapist is alone, interruption of the therapy is the most likely outcome of the conflict. Conversely, if multiple therapists are engaged, because conflict is unlikely to involve all of them simultaneously, conflict can be managed without interrupting the therapy, eventually improving the treatment outcome. The ECMBP offers a mathematical interpretation of the interpersonal cycles experienced by the BP, which might be engaged also during therapy and be a source of conflict. By highlighting a dramatic propensity towards unstable interactions, this picture is consistent with the benefits of engaging multiple therapists in the intervention to manage conflict effectively and prevent interruptions.

Our proposal builds upon prior work adopting computational modelling to explain mental illness. Within this literature, several approaches have been proposed; it is important to emphasise which specific approach is followed by the ECMBP. One approach relies on assessing behaviour of healthy individuals and patients against optimal performance, and on claiming that healthy individuals get closer to optimality than patients (Redish & Gordon, 2016). Another common approach assumes that cognitive processes of both healthy individuals and patients can be treated as optimal, for example as expression of Bayesian inference (e.g., Fletcher & Frith, 2009; Friston et al., 2014; Powers et al., 2017); the ECMBP follows the latter approach. Another distinction in the literature is between models relying on decision-making (or reinforcement learning) (e.g., Huys et al., 2015), claiming that the key features of psychopathology emerge specifically in contexts where choice and reward are engaged, and models relying on more general processes such as inference, claiming that key features of

psychopathology emerge also in contexts where no choice and reward are involved (e.g., Fletcher & Frith, 2009; Friston et al., 2014). Based on empirical evidence showing that key aspects of BP emerge even outside decision-making (e.g., Neustadter et al., 2019), the ECMBP focuses on inference processes and not solely on choice (or reward) problems. Note that the ECMBP is agnostic about the precise mechanisms adopted by the brain to perform Bayesian inference. With this regard, the literature has considered several alternative processes such as predictive coding (based on variational approaches) or sampling methods (Rabinovich et al., 2012); inference in the ECMBP can be implemented adopting either method, and the key insights offered by the model remains unaffected by which method is chosen.

In conclusion, the ECMBP advocates a simple process as being at the core of BP, namely an excessive weight attributed to cues. This model integrates traditional perspectives on BP with recent research adopting a computational approach (Fineberg, et al., 2017; 2018; Franzen et al., 2011; Henco et al., 2020; King-Casas et al., 2008; Siegel et al., 2020; Unoka et al., 2009). By investigate how patients perform judgement and decision-making tasks, computational research has offered tremendous insight but has also remained somewhat confined to rather specific domains. The ECMBP attempts to broaden the perspective and to offer a computational framework to interpret BP during everyday life, as it unfolds in ecological contexts.

References

- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (5th ed.)*. Arlington, VA: Author.
- Ball, J. S., & Links, P. S. (2009). Borderline personality disorder and childhood trauma: evidence for a causal relationship. *Current psychiatry reports, 11*(1), 63-68.
- Bishop, C. M. (2006). Machine learning and pattern recognition. *Information science and statistics*. Springer, Heidelberg.

- Bohus, M., Haaf, B., Simms, T., Limberger, M. F., Schmahl, C., Unckel, C., ... & Linehan, M. M. (2004). Effectiveness of inpatient dialectical behavioral therapy for borderline personality disorder: a controlled trial. *Behaviour research and therapy*, 42(5), 487-499.
- Borsboom, D. (2017). A network theory of mental disorders. *World psychiatry*, 16(1), 5-13.
- Bretherton, I., & Munholland, K. A. (2008). Internal working models in attachment relationships: Elaborating a central construct in attachment theory. In J. Cassidy & P. R. Shaver (Eds.), *Handbook of attachment: Theory, research, and clinical applications* (pp. 102–127). The Guilford Press.
- Burger, J., Robinaugh, D. J., Quax, R., Riese, H., Schoevers, R. A., & Epskamp, S. (2020). Bridging the gap between complexity science and clinical practice by formalizing idiographic theories: a computational model of functional analysis. *BMC medicine*, 18(1), 1-18.
- Clark, L. A., Nuzum, H., & Ro, E. (2018). Manifestations of personality impairment severity: comorbidity, course/prognosis, psychosocial dysfunction, and ‘borderline’ personality features. *Current opinion in psychology*, 21, 117-121.
- Crowell, S. E., Beauchaine, T. P., & Linehan, M. M. (2009). A biosocial developmental model of borderline personality: Elaborating and extending linehan’s theory. *Psychological bulletin*, 135(3), 495.
- D’Agostino, A., Monti, M. R., & Starcevic, V. (2019). Psychotic symptoms in borderline personality disorder: an update. *Current opinion in psychiatry*, 32(1), 22-26.
- Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: computational and mathematical modeling of neural systems*. Computational Neuroscience Series.
- Dimaggio, G., Semerari, A., Carcione, A., Nicolò, G., & Procacci, M. (2007). *Psychotherapy of personality disorders: Metacognition, states of mind and interpersonal cycles*. Routledge.
- Fineberg, S. K., Stahl, D. S., & Corlett, P. R. (2017). Computational psychiatry in borderline personality disorder. *Current behavioral neuroscience reports*, 4(1), 31-40.
- Fineberg, S. K., Leavitt, J., Stahl, D. S., Kronemer, S., Landry, C. D., Alexander-Bloch, A., ... & Corlett, P. R. (2018). Differential valuation and learning from social and nonsocial cues in borderline personality disorder. *Biological psychiatry*, 84(11), 838-845.

- Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, 10(1), 48-58.
- Fonagy, P., & Bateman, A. W. (2007). Mentalizing and borderline personality disorder. *Journal of Mental Health*, 16(1), 83-101.
- Fonagy, P., & Luyten, P. (2009). A developmental, mentalization-based approach to the understanding and treatment of borderline personality disorder. *Development and psychopathology*, 21(4), 1355-1381.
- Fonagy, P., Target, M., & Gergely, G. (2000). Attachment and borderline personality disorder: A theory and some evidence. *Psychiatric Clinics*, 23(1), 103-122.
- Fossati, A., Madeddu, F., & Maffei, C. (1999). Borderline personality disorder and childhood sexual abuse: a meta-analytic study. *Journal of personality disorders*, 13(3), 268-280.
- Franzen, N., Hagenhoff, M., Baer, N., Schmidt, A., Mier, D., Sammer, G., ... & Lis, S. (2011). Superior 'theory of mind' in borderline personality disorder: an analysis of interaction behavior in a virtual trust game. *Psychiatry research*, 187(1-2), 224-233.
- 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.
- Gluschkoff, K., Jokela, M., & Rosenström, T. (2021). General psychopathology factor and borderline personality disorder: Evidence for substantial overlap from two nationally representative surveys of US adults. *Personality Disorders: Theory, Research, and Treatment*, 12(1), 86.
- Grant, B. F., Chou, S. P., Goldstein, R. B., Huang, B., Stinson, F. S., Saha, T. D., ... & Pickering, R. P. (2008). Prevalence, correlates, disability, and comorbidity of DSM-IV borderline personality disorder: results from the Wave 2 National Epidemiologic Survey on Alcohol and Related Conditions. *The Journal of clinical psychiatry*, 69(4), 0-0.
- Hallquist, M. N., Hall, N. T., Schreiber, A. M., & Dombrovski, A. Y. (2018). Interpersonal dysfunction in borderline personality: a decision neuroscience perspective. *Current opinion in psychology*, 21, 94-104.
- Henco, L., Diaconescu, A. O., Lahnakoski, J. M., Brandi, M. L., Hörmann, S., Hennings, J., ... & Mathys, C. (2020). Aberrant computational mechanisms of social learning and decision-

- making in schizophrenia and borderline personality disorder. *PLoS computational biology*, 16(9), e1008162.
- Hopwood, C. J. (2018). Interpersonal dynamics in personality and personality disorders. *European Journal of Personality*, 32(5), 499-524.
- Huys, Q. J., Daw, N. D., & Dayan, P. (2015). Depression: a decision-theoretic analysis. *Annual review of neuroscience*, 38, 1-23.
- Huys, Q. J., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature neuroscience*, 19(3), 404-413.
- Jones, B., Heard, H., Startup, M., Swales, M., Williams, J. M. G., & Jones, R. S. P. (1999). Autobiographical memory and dissociation in borderline personality disorder. *Psychological medicine*, 29(6), 1397-1404.
- Kendzierski, D. (1980). Self-schemata and scripts: The recall of self-referent and scriptal information. *Personality and Social Psychology Bulletin*, 6(1), 23-29.
- Kernberg, O. (1967). Borderline personality organization. *Journal of the American psychoanalytic Association*, 15(3), 641-685.
- King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., & Montague, P. R. (2008). The rupture and repair of cooperation in borderline personality disorder. *science*, 321(5890), 806-810.
- Lerner, J. S., & Keltner, D. (2001). Fear, anger, and risk. *Journal of personality and social psychology*, 81(1), 146.
- Lieb, K., Zanarini, M. C., Schmahl, C., Linehan, M. M., & Bohus, M. (2004). Borderline personality disorder. *The Lancet*, 364(9432), 453-461.
- Linehan, M. M. (1993). *Cognitive-behavioral treatment of borderline personality disorder*. Guilford Publications.
- Liotti, G. (2002). The inner schema of borderline states and its correction during psychotherapy: A cognitive-evolutionary approach. *Journal of Cognitive Psychotherapy*, 16(3), 349-366.
- Liotti, G. (2014). Disorganized attachment, models of borderline states and evolutionary psychotherapy. In *Genes on the couch* (pp. 242-266). Routledge.

- Markus, H. (1977). Self-schemata and processing information about the self. *Journal of personality and social psychology*, 35(2), 63.
- Miskewicz, K., Fleeson, W., Arnold, E. M., Law, M. K., Mneimne, M., & Furr, R. M. (2015). A contingency-oriented approach to understanding borderline personality disorder: Situational triggers and symptoms. *Journal of personality disorders*, 29(4), 486-502.
- Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in cognitive sciences*, 16(1), 72-80.
- Müller, S., Wendt, L. P., & Zimmermann, J. (2021). Development and Validation of the Certainty About Mental States Questionnaire (CAMSQ): A Self-Report Measure of Mentalizing Oneself and Others. *PsyArXiv*
- Németh, N., Mátrai, P., Hegyi, P., Czéh, B., Czopf, L., Hussain, A., ... & Simon, M. (2018). Theory of mind disturbances in borderline personality disorder: A meta-analysis. *Psychiatry Research*, 270, 143-153.
- Neustadter, E. S., Fineberg, S. K., Leavitt, J., Carr, M. M., & Corlett, P. R. (2019). Induced illusory body ownership in borderline personality disorder. *Neuroscience of consciousness*, 2019(1), niz017.
- Nigg, J. T., Lohr, N. E., Westen, D., Gold, L. J., & Silk, K. R. (1992). Malevolent object representations in borderline personality disorder and major depression. *Journal of Abnormal Psychology*, 101(1), 61.
- Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.
- Pearse, L. J., Dibben, C., Ziauddeen, H., Denman, C., & McKenna, P. J. (2014). A study of psychotic symptoms in borderline personality disorder. *The Journal of nervous and mental disease*, 202(5), 368-371.
- Powers, A. R., Mathys, C., & Corlett, P. R. (2017). Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. *Science*, 357(6351), 596-600.
- Rabinovich, M. I., Friston, K. J., & Varona, P. (Eds.). (2012). *Principles of brain dynamics: global state interactions*. MIT Press.

- Redish, A. D., & Gordon, J. A. (Eds.). (2016). *Computational psychiatry: New perspectives on mental illness*. MIT Press.
- Rigoli, F. (2021). The psychology of ultimate values: A computational perspective. *Journal for the Theory of Social Behaviour*.
- Sadikaj, G., Moskowitz, D. S., Russell, J. J., Zuroff, D. C., & Paris, J. (2013). Quarrelsome behavior in borderline personality disorder: influence of behavioral and affective reactivity to perceptions of others. *Journal of abnormal psychology, 122*(1), 195.
- Schilling, L., Wingenfeld, K., Löwe, B., Moritz, S., Terfehr, K., Köther, U., & Spitzer, C. (2012). Normal mind-reading capacity but higher response confidence in borderline personality disorder patients. *Psychiatry and Clinical Neurosciences, 66*(4), 322-327.
- Semerari, A., & Fiore, D. (2007). Borderline personality disorder: model and treatment. In *Psychotherapy of personality disorders* (pp. 55-88). Routledge.
- Semerari, A., Carcione, A., Dimaggio, G., Nicolo, G., Pedone, R., & Procacci, M. (2005). Metarepresentative functions in borderline personality disorder. *Journal of personality disorders, 19*(6), 690-710.
- Sharp, C., Wright, A. G., Fowler, J. C., Frueh, B. C., Allen, J. G., Oldham, J., & Clark, L. A. (2015). The structure of personality pathology: Both general ('g') and specific ('s') factors?. *Journal of abnormal psychology, 124*(2), 387.
- Siegel, J. Z., Curwell-Parry, O., Pearce, S., Saunders, K. E., & Crockett, M. J. (2020). A computational phenotype of disrupted moral inference in borderline personality disorder. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 5*(12), 1134-1141.
- Sharp, C., Pane, H., Ha, C., Venta, A., Patel, A. B., Sturek, J., & Fonagy, P. (2011). Theory of mind and emotion regulation difficulties in adolescents with borderline traits. *Journal of the American Academy of Child & Adolescent Psychiatry, 50*(6), 563-573.
- Schmideberg, M. (1959). The borderline patient. In Arieti, S. (Ed.), *American handbook of psychiatry*. New York, NY: Basic Books

- Unoka, Z., Seres, I., Aspán, N., Bódi, N., & Kéri, S. (2009). Trust game reveals restricted interpersonal transactions in patients with borderline personality disorder. *Journal of personality disorders, 23*(4), 399-409.
- Watters, C. A., Bagby, R. M., & Sellbom, M. (2019). Meta-analysis to derive an empirically based set of personality facet criteria for the alternative DSM-5 model for personality disorders. *Personality Disorders: Theory, Research, and Treatment, 10*(2), 97.
- Widom, C. S., Czaja, S. J., & Paris, J. (2009). A prospective investigation of borderline personality disorder in abused and neglected children followed up into adulthood. *Journal of personality disorders, 23*(5), 433-446.
- Winter, D., Elzinga, B., & Schmahl, C. (2014). Emotions and memory in borderline personality disorder. *Psychopathology, 47*(2), 71-85.
- Zanarini, M. C., Ruser, T., Frankenburg, F. R., & Hennen, J. (2000). The dissociative experiences of borderline patients. *Comprehensive psychiatry, 41*(3), 223-227.
- Zweig-Frank, H., Paris, J., & Guzder, J. (1994a). Dissociation in female patients with borderline and non-borderline personality disorders. *Journal of Personality Disorders, 8*(3), 203-209.
- Zweig-Frank, H., Paris, J., & Guzder, J. (1994b). Dissociation in male patients with borderline and non-borderline personality disorders. *Journal of Personality Disorders, 8*(3), 210-218.

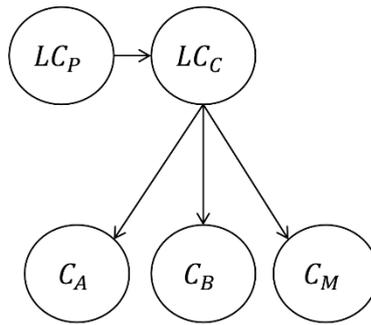


Fig 1. Graphical model proposed by the ECMBP. Variables (represented by circles) include the Past Life Context (LC_P), the Current Life Context (LC_C) and cue variables (C_A , C_B , and C_M). Arrows indicate probabilistic dependencies among variables.

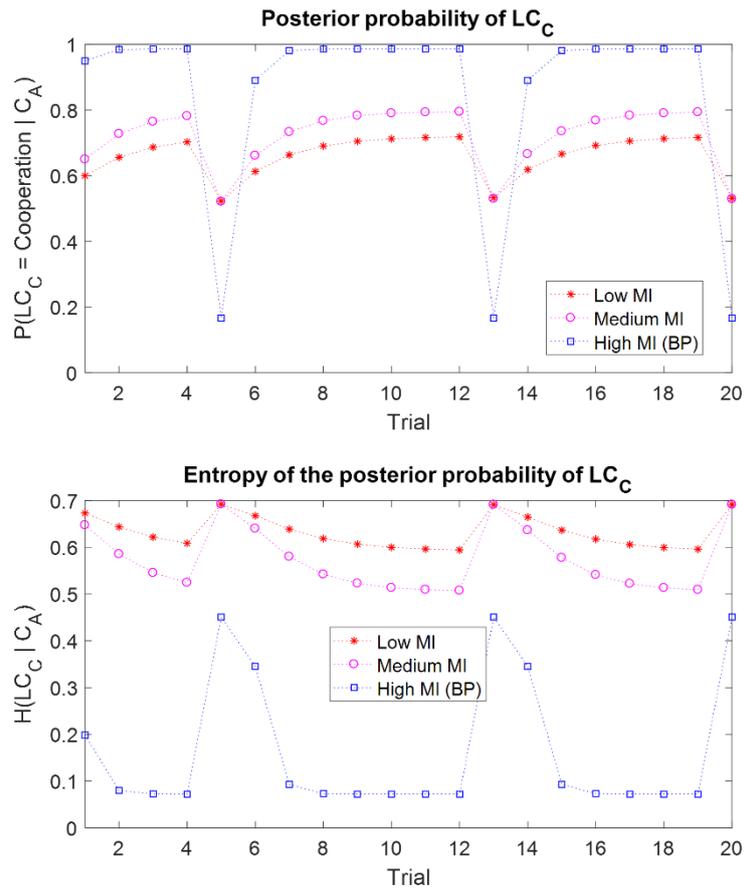


Fig. 2. Results of the first model simulation. Three agents are described, each characterised by a specific level of Mutual Information (MI) between LC_C and C_A (BP is proposed to be associated with high MI). Over trials, a Cooperation cue is always observed except for trial 5, 13, and 20, when a Betrayal cue is observed.

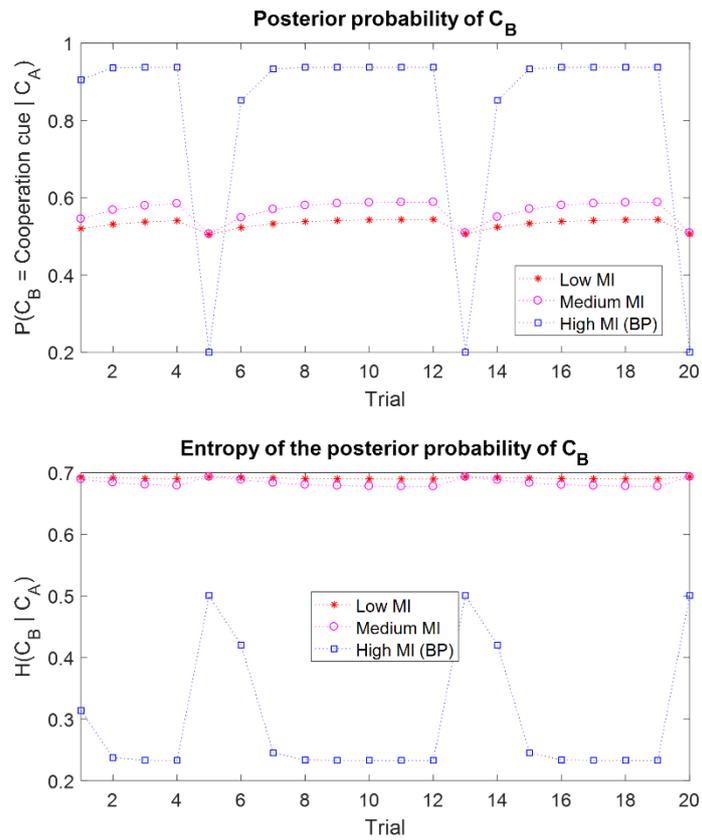


Fig. 3. Results of the second model simulation. Three agents are described, each characterised by a specific level of Mutual Information (MI) between LC_C and C_A (BP is proposed to be associated with high MI). Over trials, a Cooperation cue is always observed except for trial 5, 13, and 20, when a Betrayal cue is observed.

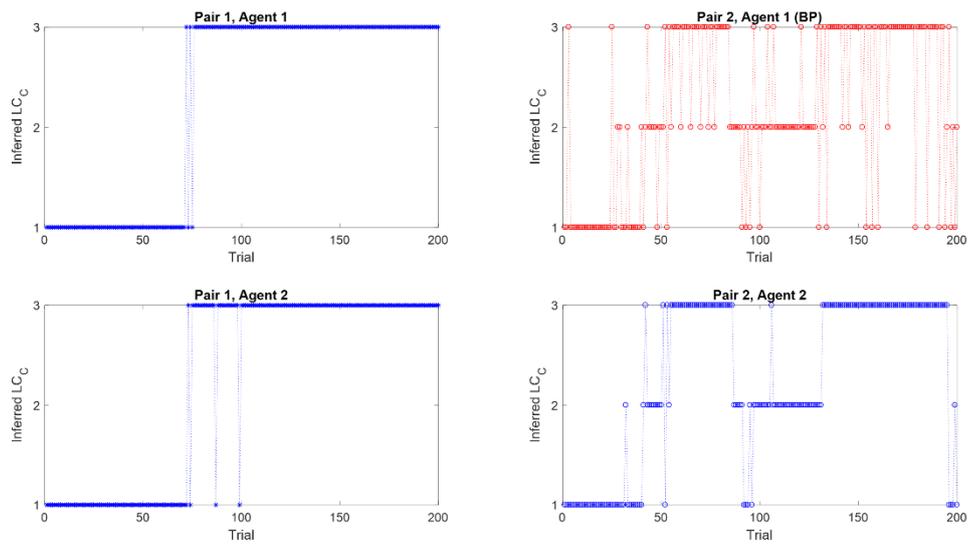


Fig. 4. Results of the fourth simulation. Two pairs of interacting agents are described. For the first pair, both agents are characterised by low Mutual Information (MI) between LC_C and C_A. For the second pair, agent two is characterised by low MI, while agent one (labelled as BP) is characterised by high MI. On every trial, the inferred LC_C is reported for each agent, corresponding to the category of LC_C associated with the highest posterior probability (1 = Cooperation; 2 = Neglect; 3 = Competition).