On the computational nature of emotions: insights from metareasoning and transformers

Jakub Onysk[1,*], Jiazhou Chen[1,†,*] and Quentin J.M. Huys[1,2]

1 Applied Computational Psychiatry Lab, Mental Health Neuroscience Department, Division of Psychiatry and Max Planck Centre for Computational Psychiatry and Ageing Research, Department for Neuroimaging, Queen Square Institute of Neurology, University College London, UK
2 North London Mental Health NHS Foundation Trust, London, UK
† Now at: Unit on Computational Decision Neuroscience (CDN), National Institute of Mental Health, Bethesda, USA
* Contributed equally

Corresponding Author
Quentin JM Huys
Max-Planck UCL Centre for Computational Psychiatry and Ageing Research
Russell Square House, 10-12 Russell Square, WC1B 5EH, London, UK
Email: q.huys@ucl.ac.uk

Word count
Abstract: 191
Total: 4908
Figures & Tables: 0
References: 173

Abstract

Emotions and affect remain enigmatic and unquestionably fundamental to our understanding of human health and illness. Affective states govern what we feel, think, say and do, but often seem to do so poorly. Why is this? This paper examines the argument that temporally-consistent emotion-like heuristics or strategies might help the brain address the fundamental metareasoning challenge in time—-what to think about and how much? This view casts established signatures of emotions in cognitive psychology as shaping the brain's internal prioritization of computations in an approximately resource-rational manner. Language and narratives support this by contributing a temporal consistency that reflects a high-level conceptual heuristic selection. Attention mechanisms and representations in transformer neural network models provide an intriguing implementation of this process and a potential avenue to study aspects of the metareasoning components of human emotions. This brings into focus a new computational view of what emotions are, why they exists and how they work, reconciling some of their objective and subjective characteristics. The paper closes with an outline of two applications, namely understanding self-report of mental-illness symptoms by quantifying the underlying affective states and understanding the neural representations of affective states.

1 Introduction

What emotions are and why they even exist remains an elusive question, despite broad agreement around their importance to individual decisions, well-being and society more broadly. One consequence of this state of affairs is that there is no formal account of emotional processes. A formal computational account could help progress by providing an overarching framework within which to integrate findings, and a path to falsification [100, 117]. One key challenge is the development of a mathematical argument outlining why something akin to emotions should exist in the first place, and why it should be necessary or helpful to brains and organisms. After all, emotions are often painted as the antipode of intellect and reason [47]. Yet, aspects of emotions appear to be present very generally across the animal kingdom. Simple learning mechanisms important in human affect are evident even in organisms as simple as aplysia [118, 119].

While there is little consensus on why emotions are necessary, there is consensus that the brain solves difficult computational problems. One critical computational challenge the brain has to address is the vast computational cost of solving any remotely relevant problem. Everyday problems such as drinking your coffee are so complex that straightforward optimal inference is prohibitively costly. The brain hence cannot afford to be optimal, but must identify approximations within the confines of its computational powers [142]. Maybe even more critically, the brain must address the metareasoning problem: how to apportion its limited computational resources. This is a decision problem about decision problems, and is radically intractable, mandating approximations [60, 127].

The core argument considered in this paper is whether emotions might serve as computational heuristics [152] enabling the brain to address the radical intractability of metareasoning problems in a manner that is approximately optimal in certain settings [55], i.e. approximately resource-rational [91]. In this view, emotions do not primarily shape behaviour, attention, specific behaviour or function as a source of independent information [31], but instead represent the brain's internal mechanisms to prioritize computations in a wholistic and temporally consistent manner to facilitate adaptive resource-rational-like behaviour in specific settings (c.f. [71]). Affective states, in this view, then become essentially computational states.

We first outline some key findings regarding the impact of emotions on perception, action and learning. Next, we introduce the metareasoning problem and ideas around bounded rationality and resource rationality to build the case for emotions as a set of approximately optimal metareasoning heuristics. Next, we consider language and its relationship to both emotions and metareasoning. We then discuss four aspects of the transformer architecture, which appear to be meaningfully related to emotions: 1) temporal consistency of outputs; 2) feature embedding of context; 3) attention 4) meta-learned in-context learning enabling Bayesian-like concept inference. Before closing with a discussion, we outline an approach to theory-driven quantification of emotions building on these notions, and discuss two applications – understanding self-report of mental-illness symptoms and the neural representations of affective states.

Before continuing, we note there are many definitions of emotion and related terms. These range from behaviourally-defined action patterns to appraisals, or constructed theories. Core affect is considered the most fundamental, integrating recent signals into a feeling described along the dimensions of valence and arousal [126]. Moods, being more diffuse and lower-intensity, can persist for hours or days and can vary without an easily identifiable cause [9], but may track a slow-moving average of recent outcomes, such as reward prediction errors, to represent the general state of the environment [43]. In contrast, emotions, implementing a more complex process, are typically more intense, discrete, and intentional, often elicited by and directed toward a specific event or object [50], reflecting individual's goals [84]. We are not able to do justice to the literature on emotion-cognition interactions more widely. Below, we focus on key concepts widely discussed and relevant to the argument we wish to propose. Henceforth, for ease of discussion, we will refer to "emotion" broadly.

2 Emotions shape computations

If emotions were indeed to function as approximate strategies to prioritize computations, then emotional states should be associated with clearly altered computational functions of the brain; and conversely altering the computational functions of the brain should alter emotions. There is extensive evidence for both directions. First, affective states influence all core computational functions of the brain including perception, inference, prediction, decisions, action preparation, action and learning. Second, altering these computational functions alters the emotional state.

Emotional states strongly influence perception and attention [101, 130, 135]. Attentional alterations are particularly prominent in emotion-related disorders, for instance depression [102]. Emotional states bias memory retrieval [35], interpretation [111, 141, 167] and predictions. The latter is emphasized in appraisal theories, where emotions are thought to involve a rapid evaluation of events along a relatively small number of discrete dimensions such novelty, valence, and goal conduciveness [50, 131] to in turn make predictions and guide action choice [132]. Situations or events, for instance, that are evaluated as being goal relevant yet incongruent might lead to frightful or defensive actions [107]. Constructionist theories argue that the subjective experience of emotion is itself an inference: the brain constructs a specific emotion concept to explain the current pattern of sensory inputs and physiological changes based on prior learning [4]. This echoes the Jamesian perspective that the subjective feeling is a recognition of the state rather than its antecedent [76].

Emotional states influence decision-making (e.g. [85, 95, 137]; see [86] for a detailed review), and recent work has extended this to more fine-grained assessments of how temporal variation in emotional states is closely related to temporal changes in decision-making [61, 129] and learning [46, 58, 159]. Emotions interface extensively with neural action systems to prioritize action intentions [6, 16, 41, 41] and prepare the body physiologically [17, 25, 27]. Affective states appear to effectively narrow the action space to a smaller subset. Evolutionary perspectives suggest this may reflect adaptations for solving recurring, high-stakes situations without the need for exhaustive deliberation [32]. Frijda reaffirmed this connection, describing affective states as flexible states of ”action readiness”—a biological bias that favors specific interactions with the world [50]. This constraint is biologically instantiated in the autonomic nervous system, which mobilizes energy to support these prioritized action tendencies. For instance, electrodermal activity tracks sympathetic arousal to reflect the body's mobilization [15], while cardiovascular measures like heart rate variability reflect the interplay between resource mobilization and self-regulation [173].

The opposite is also true: emotions not only affect perception, action and learning, but are affected by them. Causal manipulations, particularly in the setting of psychotherapy for depression and anxiety disorders, have shown the bidirectional nature of the relationship. Altering any of these computational processes reciprocally alters the emotional state. In panic disorder, for instance, altering attentional focus is a key treatment intervention [8]: patients are taught to move their attentional focus from the body (e.g. the worrying heart palpitations) to their environment (e.g. examining the colours around them). Cognitive therapies involve changing thoughts, interpretations, and challenging predictions [7]. Metacognitive therapies extend these changes to higher-level motivations [160]. In dialectic behaviour therapy [94], controlling emotions is achieved by acting 'against' the emotion. For instance, when angry with a person, anger can be reduced by considering what presents might bring that same person pleasure. Hence, altering the action evaluation set has immediate influences on the subjective emotional state. These effects are in keeping with a prioritization of specific computations. Altering attention alters which stimuli are prioritized for processing. Altering action plans alters which actions are evaluated. Altering thoughts directly alters the interpretations and predictions the brain evaluates. More broadly, many psychotherapeutic interventions attempt to alter what people “choose” to think of in emotionally relevant moments.

Overall, then, emotional states are intricately and causally interwoven with the core computations the brain engages in. Emotional states are altered by computations (e.g. learning, inference, actions or action plans), but also directly shape computations (e.g. biased interpretation, memory retrieval, decisions pol-

icy). This mutually causal relationship suggests that emotional states may play an important, inherently computational role, prioritizing certain computations over others.

3 Emotion as metareasoning heuristics

Why would the brain require approximate strategies to prioritize computations? Formally, the problem of prioritizing computations ("choosing what to think about") is the metareasoning problem. It arises when computations are costly and budget is limited, so that computational problems cannot be evaluated in their completeness [60, 91, 127, 128]. Consider playing chess. If there are around 30 options at each move, then thinking ahead n moves requires evaluation of 30^n combinations of unique sequence of moves. The problem is even bigger when you consider that the brain has to decide whether or not to evaluate each option at each move, which blows up the evaluation space to the astronomical $2^{(30+30^1+30^2+...+30^n)}$. Clearly this is not possible.

Humans and animals employ heuristic computational strategies to reduce the metareasoning problem by prioritizing a restricted set of options or computations. In chess, beginners only evaluate a tiny number of options, and only look one or two steps ahead at best. Masters look deep into the tree, but do so very selectively. Selecting which components of the decision tree to focus on is hard. Generally, humans utilize a number of strategies to simplify the problem, such as pruning [69, 70], memoization and subgoal substitution [29, 69], feature learning [154], risk optimization [51, 138] and replay [104] amongst others. This prioritization often goes awry in mental illnesses: during a state of craving in addictive behaviour, the brain prioritizes thinking about the advantages of drug taking but not other rewarding experiences; when worrying, the evaluation of aversive outcomes is prioritized over the positive ones; in depression, rumination involves prioritizing thinking about perceived failures over productive review of past successes. Hence, for heuristics to meaningfully address the metareasoning problem adaptively, the heuristic selection itself needs to be relatively straightforward and robust, and hence either the number of discrete heuristics relatively small or the space of heuristics relatively simple. At the same time, a substantial degree of flexibility needs to be retained, allowing, for instance, escape behaviours to be sensitive to the direction of the predator or the closeness of a potential safety space [39, 40, 71, 74, 168].

Heuristics that alter metareasoning will have broad effects on many of the brain's observable outputs. While the case for the basic cognition (e.g. perception, inference and action [20]) is maybe straightforward, the effect on higher order cognition is more intricate. For example, the heuristics might affect learning and planning by changing expectations [21, 23]. Depending on what future options were predicted, an outcome may or may not be surprising, which further influences future learning and behaviour [33]. Anxiety disorders present one such interesting case where the metareasoning strategies influence perception, action and learning in coordinated and consistent manner. Extensive evaluation of aversive outcomes drives avoidance, which in turn can prevent learning [108, 122]. Such correlated changes might be highly adaptive in certain situations such as highly unpredictable environments or wholly-negative environments, but often lead to sub-optimal behaviour.

The metareasoning problem therefore implies the need for computational heuristics: the brain requires inexpensive ways of deciding which computations to focus under limited resources. For such heuristics to be useful, they must be simple enough to select, yet broad enough to influence multiple cognitive domains. As reviewed above, emotions are automatic and guide interpretation, memory retrieval, action preparation, language and learning in a coordinated manner: they alter which features are sampled, which memories are recalled, which actions enter the consideration set, and which outcomes warrant further evaluation. Therefore, emotional states themselves may be heuristics for solving the metareasoning problem. They are context-sensitive but generalizable: a state such as anxiety, anger or sadness does not prescribe a single action, but sets a pattern of computational priorities that can be reused across situations with shared structure. Their usefulness does not require that they improve every local decision.

Another reason why emotional states are plausible heuristics for the metareasoning problem is that the

world itself is temporally structured. Emotional states have temporal persistence, sometimes described as momentum [11, 42, 45]. Emotional phenomena vary in this feature, with emotions seen as most fleeting, moods as somewhat longer, and personality characteristics as very long-term. This persistence matters because many ecologically relevant problems unfold over sequences rather than isolated choices. When the broader situation remains similar, a continuing affective state can preserve a pattern of priorities across successive computations: attention remains tuned to certain features, memory retrieval favours certain episodes, action preparation stays oriented toward certain goals, and learning signals are interpreted against the same context. Indeed, emotional states are characterized by an urge to persist with a certain thought or action or inference: When worrying, it is very difficult to focus on the potential positive outcomes. People who ruminate struggle to identify the intrinsic causes for their successes. When craving drugs, it is difficult to find enjoyment in more pedestrian ambulation. When angry, it is difficult to think of pleasant surprises for one's perceived opponent. While this is recognized by psychotherapeutic interventions, which often aim to correct this maladaptive prioritization [160], this temporal dimension is often underemphasized in cognitive science, where perception and decision-making are commonly studied one choice at a time. When sequences are considered, there has typically been a focus on short temporal windows, such as in Markov decision problems, possibly with some latent state inference [54, 103, 133]. In this sense, emotion supports sequences of locally reasonable computations without requiring exhaustive replanning at every step.

4 Language, emotions and computations

How does language relate to metareasoning and emotions? Arguably, language occupies a pivotal role at the nexus of emotion, metareasoning and computation.

First, language is critical to human emotion assessment. Emotional states are usually measured using self-reported instruments that rely on language. The PANAS scale [158], for instance, asks participants to judge to what extent individual emotion words such as interested, excited or hostile describe how they have felt over the past week. Self-report scales such as the Patient Health Questionnaire-9 similarly ask patients to judge symptoms presence over an extended recent period [82], while observer-reported scales rely on an observer to make this assessment through dialogue and observation. This not only requires introspective (Serfaty and Huys, in prep.), mnemonic and retrospective valuation [10], but also taps into individual semantic representations of emotion terms, which change over development [66]. Indeed, the semantic structure of self-report assessments accounts for a substantial fraction of the apparent interrelationship between assessments [162]. Furthermore, language output contains quantifiable information about self-reported emotional states [80, 96, 115, 136], suggesting that emotional states consistently relate to the internal computational states determining language production.

Second, the metareasoning challenges encountered above are also apparent in language production. More obviously, the problem takes a specifically linguistic form: selecting lexical items, assembling grammatical structure, maintaining discourse coherence and carrying affective tone. Akin to decision-making or action scenarios, the computational cost of evaluating every possible phrasing to find the optimal one is ruinous. However, language production is also a process that extends far beyond the simple production of words and incorporates many aspects of action and planning. "The utterances people produce are crafted with great sophistication to satisfy multiple goals at different communicative levels. For example, in a single utterance, a speaker may inform a hearer of two or more propositions, make a request, shift the focus of the discourse, and flatter the hearer." [3]. As such, language is necessarily exquisitely sensitive to the metareasoning demands identified above.

Third, the combination of the sensitivity of language to metareasoning challenges, and its success in practice, suggests that the metareasoning problem is being constrained. The relevant constraints are not only grammatical, they also reflect what is currently salient, uncertain, threatening, rewarding or action-relevant. This parallels the role proposed for emotional states more generally. If emotional states function as metareasoning heuristics, then language production is one domain in which they govern access and

priority, alongside interpretation, memory, action preparation and learning. Thus, the language people produce is not separate from the underlying emotional state: the state helps give rise to the descriptions and those descriptions can then carry the state forward or revise it. A statement such as ”the talk went terribly; I am worthless” turns a high-dimensional episode into a compact semantic focus, making some later interpretations easier to retrieve and others less likely to be considered. Indeed, states induced through language perception or production influence a broad range of cognitive computations, including perception [97], decision-making [26, 30], short-term memory [115], action and learning. In experimental settings, this is reflected in the importance of instructions. In psychiatry, language delivered in the form of psychotherapeutic interventions has dissociable and specific effects on decision-making [112, 113]. Many such interventions can be understood as attempts to instil different metareasoning priors by offering alternative ways of summarising behavioural options, evaluations and goals through language.

These points strongly suggest that the computational states determining language might be intricately related to the computational states underlying perception, action and learning, and that metareasoning approximations might apply across these domains. Specifically, language induces consideration sets that constrain the options considered not only in language production, but also in planning, inference, judgement and learning; these impose functional constraints akin to metareasoning heuristics; and conversely language production—notably in emotion judgements—is similarly affected by consideration sets induced by action plans, perception and inference. Language is therefore a particularly useful domain for understanding emotions as metareasoning heuristics in two ways. First, language is itself a domain in which the metareasoning problem is readily resolved: during comprehension and production, the language system compresses high-dimensional reality and a vast space of possible meanings, word choices, grammatical structures and pragmatic interpretations into a coherent sequence of words. Studying this reduction may help identify mechanisms for constraining high-dimensional cognitive possibilities, the same functional role proposed above for emotional states. Second, language is closely tied to emotion: it is used to report and induce emotional states, and its production is shaped by them. It therefore provides a domain in which emotional relevance and metareasoning constraints can be studied together.

Overall, language is closely related to emotion perception and judgement, can induce and change emotional states, affects perceptions, decisions, planning and learning in a manner comparable to emotions, and has computational demands rendering it similarly sensitive to metareasoning heuristics. We next consider how advances in the modelling of language might provide insights into the nature of emotional states.

5 Language transformers

If human language production shares this metareasoning problem yet consistently yields functional communication and cognitive regulation, then it may be fruitful to consider principles in language transformer systems that have yielded one possible solution.

The advanced large language models (LLMs) that have revolutionized language production and analysis are built on the transformer architecture [155]. These transformers capture human language production and representation in some detail. They are able to consistently generate fluent language and converse in response to a set of prompts and how humans communicate, think, and report mental states using language [28, 38, 57, 153]. They share with humans aspects of syntactic knowledge acquisition [37, 170] capture human language prediction [56], and likely many other dimensions of human language processing [98, 149]. Internal representations of LLMs show some relationship to human neural activity during language generation, story listening and internal speech generation [56, 147, 148, 150]. They also perform well on emotional inference tasks [53, 143, 169], showing a significant extent of conceptual alignment [88]. Notably, they also perform well on mental-health tasks [166], offering a way to model depression symptoms [48, 67, 106, 115] supplement mental state assessment [12, 145], and other assessments tasks [64, 161]. LLM-derived features can be predictive of symptom improvement in

treatment settings [1], with LLM applications in therapy settings showing promise [49, 92, 109, 110], although with some important safety caveats [24, 65, 72, 134, 144]. Indeed, the generation of consistent responses across contexts has sparked a growing interest and success in using LLMs to model human behaviour and cognition across hundreds of tasks [13]. This clearly establishes LLMs' ability to exploit semantic representations to capture internal states that are pivotal to solving a behavioural or cognitive problem [75].

In displaying this ability to integrate vast contexts into behaviourally, cognitively and neurally relevant states, LLMs function analogously to how we conceptualised emotions – efficient, resource-optimal heuristics for guiding behaviour. Indeed, transformer LLMs remain sensitive to inference costs and can be further improved by explicitly considering the value of computation [34]. However, their performance already ensures they address a substantial fraction of the metareasoning problems outlined above across domains. How do they achieve this, and can this help provide insights into the nature of emotions?

Fuelling this LLM capability is a deep neural network, consisting of many computational layers that process the input words (or, more precisely 'tokens') as it passes through the network predicting the next word (token). To achieve this, transformers accept a representation of inputs in terms of features (an embedding), and transform this iteratively until a new representation is achieved which can be linearly decoded to reliably identify the next word in the sequence [151]. Importantly, the architecture of LLMs allows for rich representations of very long contexts [18, 155], constraining the space of possible next words, which are consistent with a given context, maximising predictive performance.

At the heart of transformer LLMs are two transformations: the so-called self-attention mechanism, whereby the currently considered token is compared to every past token in the sequence; and the key-value representation whereby the token similarities are represented separately from the token contents [73, 121, 124]. The predictions from each past token are then weighted by this similarity. Transformers have multiple such similarity comparisons in parallel (so-called multi-head attention), and repeat this multiple times over multiple layers.

This structure leads to striking emergent abilities. The most relevant one to the current discussion is probably in-context learning [99], whereby the model can be given 'instructions' as a verbal prompt, and then perform, without further updating the model itself, a variety of 'tasks'. In-context learning has been shown to rely on Bayesian concept inference [2], whereby the prompts or instructions induce a narrow, focused distribution over latent concepts [164, 165, 171]. This narrow 'selection' of latent concepts then leads to weighting potential future tokens in relation to the relevant concept, which has also been cast as resource-rational inference [163]. Latent concept selection can be seen as a generalized version of induction seen in smaller networks [44, 114], where some attention heads select (via attention) those past tokens that look most similar to the current token and lead to the prediction that the same continuation will follow. What is striking about the scaling of this process in LLMs are arguments that it leads to very robust harnessing of multi-word semantics [19], yet what emerges is a relatively simple linear representation [77].

In-context learning in fact emerges from a broader process of meta-learning that enables contextual adaptation ranging from basic language capabilities to flexible use of new information for a task at hand [83]. Meta-learning describes how during pre-training, transformers are exposed to diverse range of contexts that rely on a mixture of recurring latent processes ranging from simple (e.g. storage and retrieval) to complex tasks (such as analogical reasoning or instruction following). In-context learning relies on these latent processes to produce relevant classes of outputs that can solve novel instantiations of a task. As these processes are not explicitly defined, the meta-learning must take place to establish what is to be learned and how, and in which context is that latent process relevant [71, 79, 116]. Indeed, the prioritization of evaluation can be though of as being meta-learned based on the history of rewards, actions and states to be then deployed in context [22].

Emotions can now be considered meta-learned heuristic strategies that arise in-context – for example biased thought trajectories (worry, rumination) in specific situations. Specifically, meta-learned input-output tendencies would later be re-utilised in "affectively" determined in-context settings. One example

is the ability of LLMs to use in-context learning to rewrite passages in different emotional tones and in their ability to representat conceptual dimensions in psychopathology [115]. To achieve this, in-context learning relies on accurate emotion representations as concepts [125] that, strikingly, appear to be relatively simple and linear [123, 169]. Hence, the parallel between emotions and transformers steeped in the framework of metareasoning and meta-learning could intriguingly provide a plausible mechanistic approach, justifying the use of LLMs in the quantification of emotional states. Furthermore, these findings echoes findings in human and animal meta-learning in the prefrontal cortex [22, 146, 156] and its interaction with other brain regions involved in reward preprocessing, like ventral tegmental area or striatum [62, 105].

Coming back to the main argument, these algorithmic advances in transformer models of language suggest a) that aspects of metareasoning that emotions solve, could be solved by a transformer-like architecture via similarly meta-learned strategies, possibly even in their neural circuits; b) that emotions might be closely aligned with the attentional (in the sense of transformers) selection of relevant predictors to guide processing in approximately Bayesian way; and c) that the emergent representation of emotional concepts might indeed be relatively simple (linear!).

6 Discussion

Emotions and brain computations are causally mutually related. Emotions causally alter all key higher cognitive functions of the brain: inference, action and learning. Emotions are also causally altered by these functions – changing perception, action plans or learning changes the emotional states. Brains face a key challenge in performing computations related to inference, action and learning: that of appropriately apportioning resources. This resource allocation problem (metareasoning) is computational intractable, and demands approximations. We have outlined how heuristic approximations might have similarities to emotions in terms of their relationship to inference, action and learning. We then considered language, outlining how it is intricately related to inference, action, learning and emotional states; and shares computational demand characteristics. This finally led us to consider how advances in the modelling and analysis of language through transformer models might yield insights into how neural-like structures solve the metareasoning problem and develop emotions as correlates of concepts in transformers. Finally, emerging research on representation of latent (emotion) concepts in transformers suggests the postulated simplicity of representation facilitating computational expediency.

LLMs have been successfully used to decode emotion labels from various data [36, 48, 78, 106, 139, 166]. These efforts relied on ground-truth labelling and then the use of supervised techniques. By contrast, here, we have attempted to relate the structure of transformers to the hypothesized structure and function of emotions as metareasoning approximations. How exactly this can be put to use remains to be seen. However, it may allow LLM architectures to be used to directly probe emotion processes [13]. It also suggests further exploration of neural correlates with components of LLM state representations [56, 67, 147]. One critical test will be whether such decoders can successfully recover subjectively reported affective states. We discuss these notions in the Clinical Implications section.

The central thesis in this paper invites a re-framing of complex psychological constructs such as emotions within a computationally tractable framework. We have done so by considering the relationship between emotion, computational functions of the brain, metareasoning, language and transformer models. This view then casts affective states as computational states that: 1) integrate information about past experiences; 2) maximise the temporal and contextual state consistency by narrowly constraining the consideration set; 3) respond to and integrate relevant contextual changes to support adaptive behaviour. We argued that this characterisation, due to its functional parallels, is particularly amenable to quantification with transformer-based LLMs. As such, the contribution of this paper is to consider how computational considerations can be used to steer the study of subjective experience from a qualitative perspective towards a quantitative one.

7 Future Research Directions and Clinical Implications

The views outlined make a number of specific predictions which need to be tested. Two potential applications, appear particularly interesting: the study of psychopathology self-reports; and the study of neural representations of affective computational states.

Self-reports show important correlations and structure, and it has traditionally been assumed that these factors relate to an underlying affective state. LLMs capture self-report covariance in interesting detail [52, 162], suggesting a potential relationship between such factors and latent LLM representations. A recent study obtained open-ended responses to questionnaires as a way of sampling from the internal computational state, in addition to standard self-report [67, 115]. This pair of ground-truth observations enables a kind of evaluation of how LLM internal representations relate to self-rater report. It also enables causal manipulations through the induction of affective states via langauge stimuli and suggested that a meaningful relationship between internal LLM states and the factors underlying self-report might exist [115]. Clinically, an interesting direction might be the identification of new features and dimensions in open-ended samples, examining how they relate to (the improvement in) symptoms [1, 67]. Investigating the interaction between such features over time in response to external inputs might enable insights into mechanisms of psychological therapy. The generative nature of LLMs could then offer a way to perform in-silico experiments assessing the effectiveness of interventions given a specific affective state, potentially aiding the design of new interventions. As such and given the growing industry of mental health chatbots and AI therapists, we think it is paramount to establish and perform thorough set of evaluation criteria to ensure an accurate representation of an affective state – crucial for safe and effective intervention [1, 65, 134].

In terms of neural representations, emotions appear to represent a network of regions [5, 63, 93, 120, 140, 172], rather than "coding" in particular brain regions, i.e. a contextual construction from the interactions of domain-general brain networks that support other cognitive functions. Such a context dependency creates a generalizability problem for objectively quantifying affective states with neural signals: neural activity shows substantial variance within emotion categories even after controlling for induction methods [140] and emotion classification from neural signals seems to perform best when restricted to specific induction domains [81, 87, 89, 90, 157]. The view of emotions and LLMs outlined above suggests that LLMs may be useful in probing representations underlying such contextually sensitive affective states. One approach is to build on relating neurally evoked activity in response to auditory or visual narratives to LLM-derived semantic embeddings [56, 68, 147]. What would be particularly interesting is to see whether there exist components of neural activity derived from such embedding that predict self-reported emotion judgements across contexts, e.g. across a movie, story, and a cognitive task with momentary mood judgements [14]. If successful, such signals could potentially provide important supplement mental state assessments in research settings. With the appropriate consideration of the very substantial ethical challenges around mental privacy [147] such decoders could even be developed towards objective assessments of affective states in clinical settings, for instance to measure affective states when disclosure is challenging [59].

Author contributions

JO, JC contributed equally
JO, JC, QJMH: conception, main draft and revisions.

Funding

Part of this work was supported by a Wellcome Trust grant to QJMH (221826/Z/20/Z). QJMH was employed by University College London during this work.

JO was supported by the International Max Planck Research School on Computational Methods in Psychiatry and Ageing Research UCL (IMPRS COMP2PSYCH) fellowship (577749/D-CON/186534).

JC was supported by a UCL-NIMH studentship.

QJMH has obtained fees and options for consultancies for Aya Technologies and Alto Neuroscience.

QJMH has received research grant funding from Carigest S.A., German Research Foundation, Koa Health, NIHR, Swiss National Science Foundation, Wellcome Trust. QJMH acknowledges support by the NIHR UCLH BRC and NIHR MH-TRC MHM.

Competing Interests

QJMH has obtained fees and options for consultancies for Aya Technologies and Alto Neuroscience.

JO, JC declare no competing interests.

References

[1] **Abdou M**, **Sahi RS**, **Hull TD**, **Nook EC**, **Daw ND** (2025). Leveraging large language models to estimate clinically relevant psychological constructs in psychotherapy transcripts. *Computational Psychiatry* .

[2] **Agarwal N**, **Dalal SR**, **Misra V** (2025). The bayesian geometry of transformer attention. ArXiv:2512.22471 [cs].

[3] **Appelt D** (1982). Planning natural language referring expressions. In *20th Annual Meeting of the Association for Computational Linguistics*, pages 108–112.

[4] **Barrett LF** (2017). *How emotions are made: The secret life of the brain*. How emotions are made: The secret life of the brain. Houghton Mifflin Harcourt, Boston, MA.

[5] **Barrett LF** (2017). The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience* **12**(1), 1–23.

[6] **Barrett LF**, **Adolphs R**, **Marsella S**, **Martinez A**, **Pollak SD** (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological science in the public interest : a journal of the American Psychological Society* **20**(1), 1–68.

[7] **Beck AT** (1987). Cognitive models of depression. *J Cog Psychotherapy. Int Quart.* **1**, 5–37.

[8] **Beck AT** (2013). Cognitive approaches to panic disorder: Theory and therapy. In *Panic*, pages 91–109. Routledge.

[9] **Beedie C**, **T Peter**, , **Lane A** (2005). Distinctions between emotion and mood. *Cognition and Emotion* **19**(6), 847–878.

[10] **Ben-Zeev D**, **Young MA** (2010). Accuracy of hospitalized depressed patients' and healthy controls' retrospective symptom reports: an experience sampling study. *J Nerv Ment Dis* **198**(4), 280–285.

[11] **Bennett D**, **Davidson G**, **Niv Y** (2022). A model of mood as integrated advantage. *Psychological review* **129**, 513–541.

[12] **Bi G**, **Chen Z**, **Liu Z**, **Wang H**, **Xiao X**, **Xie Y**, **Zhang W**, **Huang Y**, **Chen Y**, **Peng L**, **Huang M** (2025). MAGI: Multi-Agent Guided Interview for Psychiatric Assessment. In W Che, J Nabende, E Shutova, MT Pilehvar, eds., *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24898–24921. Association for Computational Linguistics, Vienna, Austria.

[13] **Binz M**, **Akata E**, **Bethge M**, **Brändle F**, **Callaway F**, **Coda-Forno J**, **Dayan P**, **Demircan C**, **Eckstein MK**, **Éltető N**, **Griffiths TL**, **Haridi S**, **Jagadish AK**, **Ji-An L**, **Kipnis A**, **Kumar S**, **Ludwig T**, **Mathony M**, **Mattar M**, **Modirshanechi A**, **Nath SS**, **Peterson JC**, **Rmus M**, **Russek EM**, **Saanum T**, **Schubert JA**, **Schulze Buschoff LM**, **Singhi N**, **Sui X**, **Thalmann M**, **Theis FJ**, **Truong V**, **Udandarao V**, **Voudouris K**, **Wilson R**, **Witte K**, **Wu S**, **Wulff DU**, **Xiong H**, **Schulz E** (2025). A foundation model to predict and capture human cognition. *Nature* .

[14] **Blain B**, **Rutledge RB** (2020). Momentary subjective well-being depends on learning and not reward. *eLife* **9**, e57977.

[15] **Boucsein W** (2012). *Electrodermal Activity*. Springer US, Boston, MA.

[16] **Braine A**, **Georges F** (2023). Emotion in action: When emotions meet motor circuits. *Neuroscience and biobehavioral reviews* **155**, 105475.

[17] **Brosschot JF**, **Thayer JF** (2003). Heart rate response is longer after negative emotions than after positive emotions. *International Journal of Psychophysiology* **50**(3), 181–187.

[18] **Brown TB**, **Mann B**, **Ryder N**, **Subbiah M**, **Kaplan J**, **Dhariwal P**, **Neelakantan A**, **Shyam P**, **Sastry G**, **Askell A**, **Agarwal S**, **Herbert-Voss A**, **Krueger G**, **Henighan T**, **Child R**, **Ramesh A**, **Ziegler DM**, **Wu J**, **Winter C**, **Hesse C**, **Chen M**, **Sigler E**, **Litwin M**, **Gray S**, **Chess B**, **Clark J**, **Berner C**, **McCandlish S**, **Radford A**, **Sutskever I**, **Amodei D** (2020). Language Models are Few-Shot Learners. ArXiv:2005.14165 [cs].

[19] **Bu D**, **Han A**, **Huang W**, **Nitanda A**, **Suzuki T**, **Wong HS**, **Zhang Q** (2024). Provably transformers harness multi-concept word semantics for efficient in-context learning. In *Advances in Neural Information Processing Systems 37*, NeurIPS 2024, pages 63342–63405. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

[20] **Callaway F**, **Rangel A**, **Griffiths TL** (2021). Fixation patterns in simple choice reflect optimal information sampling. *PLOS Computational Biology* **17**(3), e1008863.

[21] **Callaway F**, **Yu M**, **Mattar MG** (2024). Revealing human planning strategies with eye-tracking. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

[22] **Chen S**, **Callaway F**, **Kumar S**, **Lupkin SM**, **Wallis JD**, **McGinty VB**, **Rich EL**, **Mattar MG** (2026). Learning to select computations in recurrent neural circuits.

[23] **Chen S**, **Jensen KT**, **Mattar MG** (2025). Rational decisions in multi-step environments with few rollouts.

[24] **Cheng M**, **Lee C**, **Khadpe P**, **Yu S**, **Han D**, **Jurafsky D** (2026). Sycophantic ai decreases prosocial intentions and promotes dependence. *Science (New York, N.Y.)* **391**, eaec8352.

[25] **Christopoulos GI**, **Uy MA**, **Yap WJ** (2016). The Body and the Brain: Measuring Skin Conductance Responses to Understand the Emotional Experience. *Organizational Research Methods* .

[26] **Circi R**, **Gatti D**, **Russo V**, **Vecchi T** (2021). The foreign language effect on decision-making: A meta-analysis. *Psychonomic Bulletin & Review* **28**(4), 1131–1141.

[27] **Clauss K**, **Gorday JY**, **Bardeen JR** (2022). Eye tracking evidence of threat-related attentional bias in anxiety- and fear-related disorders: A systematic review and meta-analysis. *Clinical Psychology Review* **93**, 102142.

[28] **Colombatto C**, **Fleming SM** (2024). Folk psychological attributions of consciousness to large language models. *Neuroscience of Consciousness* **2024**(1), niae013.

[29] **Correa CG**, **Ho MK**, **Callaway F**, **Daw ND**, **Griffiths TL** (2023). Humans decompose tasks by trading off utility and computational cost. *PLOS Computational Biology* **19**(6), e1011087.

[30] **Costa A**, **Vives M**, **Corey JD** (2017). On Language Processing Shaping Decision Making. *Current Directions in Psychological Science* **26**(2), 146–151.

[31] **Damasio A** (1994). *Descartes' error. Emotion, Reason and the Human Brain*. Putnam New York.

[32] **Darwin C** (1872). The expression of the emotions in man and animals. *Darwin. The Indelible Stamp* pages 1061–1257.

[33] **Daw ND**, **Gershman SJ**, **Seymour B**, **Dayan P**, **Dolan RJ** (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**(6), 1204–1215.

[34] **De Sabbata CN**, **Sumers TR**, **AlKhamissi B**, **Bosselut A**, **Griffiths TL** (2024). Rational metareasoning for large language models.

[35] **Dehon H**, **Larøi F**, **Van der Linden M** (2010). Affective valence influences participant's susceptibility to false memories and illusory recollection. *Emotion* **10**(5), 627–639.

[36] **Demszky D**, **Movshovitz-Attias D**, **Ko J**, **Cowen A**, **Nemade G**, **Ravi S** (2020). GoEmotions: A Dataset of Fine-Grained Emotions. ArXiv:2005.00547 [cs].

[37] **Diego-Simón P**, **D'Ascoli S**, **Chemla E**, **Lakretz Y**, **King JR** (2024). A Polar coordinate system represents syntax in large language models. *Advances in Neural Information Processing Systems* **37**, 105375–105396.

[38] **Dillion D**, **Tandon N**, **Gu Y**, **Gray K** (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences* **27**(7), 597–600.

[39] **Domenici P**, **Blagburn JM**, **Bacon JP** (2011). Animal escapology i: theoretical issues and emerging trends in escape trajectories. *The Journal of experimental biology* **214**, 2463–2473.

[40] **Domenici P**, **Blagburn JM**, **Bacon JP** (2011). Animal escapology ii: escape trajectory case studies. *The Journal of experimental biology* **214**, 2474–2494.

[41] **Ekman P**, **Friesen WV**, **O'Sullivan M**, **Chan A**, **Diacoyanni-Tarlatzis I**, **Heider K**, **Krause R**, **LeCompte WA**, **Pitcairn T**, **Ricci-Bitti PE**, **Scherer K**, **Tomita M**, **Tzavaras A** (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology* **53**(4), 712–717.

[42] **Eldar E**, **Roth C**, **Dayan P**, **Dolan RJ** (2018). Decodability of reward learning signals predicts mood fluctuations. *Current biology : CB* **28**, 1433–1439.e7.

[43] **Eldar E**, **Rutledge RB**, **Dolan RJ**, **Niv Y** (2016). Mood as Representation of Momentum. *Trends in Cognitive Sciences* **20**(1), 15–24.

[44] **Elhage N**, **Nanda N**, **Olsson C**, **Henighan T**, **Joseph N**, **Mann B**, **Askell A**, **Bai Y**, **Chen A**, **Conerly T**, **DasSarma N**, **Drain D**, **Ganguli D**, **Hatfield-Dodds Z**, **Hernandez D**, **Jones A**, **Kernion J**, **Lovitt L**, **Ndousse K**, **Amodei D**, **Brown T**, **Clark J**, **Kaplan J**, **McCandlish S**, **Olah C** (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread* Https://transformer-circuits.pub/2021/framework/index.html.

[45] **Erdman A**, **Eldar E** (2023). The computational psychopathology of emotion. *Psychopharmacology* **240**, 2231–2238.

[46] **Eshel N**, **Roiser JP** (2010). Reward and punishment processing in depression. *Biol Psychiatry* **68**(2), 118–124.

[47] **Evans JSBT** (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annu Rev Psychol* **59**, 255–278.

[48] **Farruque N**, **Goebel R**, **Sivapalan S**, **Zaiane OR** (2024). Depression symptoms modelling from social media text: an LLM driven semi-supervised learning approach. *Language Resources and Evaluation* **58**(3), 1013–1041.

[49] **Filienko D**, **Wang Y**, **Jazmi CE**, **Xie S**, **Cohen T**, **De Cock M**, **Yuwen W** (2025). Toward Large Language Models as a Therapeutic Tool: Comparing Prompting Techniques to Improve GPT-Delivered Problem-Solving Therapy. *AMIA Annual Symposium Proceedings* **2024**, 417–426.

[50] **Frijda NH** (1986). *The emotions*. The emotions. Editions de la Maison des Sciences de l'Homme, Paris, France.

[51] **Gagne C**, **Dayan P** (2021). Two steps to risk sensitivity. In M Ranzato, A Beygelzimer, Y Dauphin, P Liang, JW Vaughan, eds., *Advances in Neural Information Processing Systems*, volume 34, pages 22209–22220. Curran Associates, Inc.

[52] **Galatzer-Levy IR**, **Tomasev N**, **Chung S**, **Williams G** (2026). Generative Psychometrics–An Emerging Frontier in Mental Health Measurement. *JAMA Psychiatry* **83**(1), 5–6.

[53] **Gandhi K**, **Lynch Z**, **Fränken JP**, **Patterson K**, **Wambu S**, **Gerstenberg T**, **Ong DC**, **Goodman ND** (2024). Human-like Affective Cognition in Foundation Models. ArXiv:2409.11733 [cs].

[54] **Gershman SJ**, **Niv Y** (2012). Exploring a latent cause theory of classical conditioning. *Learn Behav* **40**(3), 255–268.

[55] **Gigerenzer G**, **Goldstein DG** (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological review* **103**, 650–669.

[56] **Goldstein A**, **Zada Z**, **Buchnik E**, **Schain M**, **Price A**, **Aubrey B**, **Nastase SA**, **Feder A**, **Emanuel D**, **Cohen A**, **Jansen A**, **Gazula H**, **Choe G**, **Rao A**, **Kim C**, **Casto C**, **Fanda L**, **Doyle W**, **Friedman D**, **Dugan P**, **Melloni L**, **Reichart R**, **Devore S**, **Flinker A**, **Hasenfratz L**, **Levy O**, **Hassidim A**, **Brenner M**, **Matias Y**, **Norman KA**, **Devinsky O**, **Hasson U** (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience* **25**(3), 369–380.

[57] **Hagendorff T**, **Dasgupta I**, **Binz M**, **Chan SCY**, **Lampinen A**, **Wang JX**, **Akata Z**, **Schulz E** (2024). Machine Psychology. ArXiv:2303.13988 [cs] version: 6.

[58] **Halahakoon DC**, **Kieslich K**, **O'Driscoll C**, **Nair A**, **Lewis G**, **Roiser JP** (2020). Reward-processing behavior in depressed participants relative to healthy volunteers. *JAMA Psychiatry* .

[59] **Hallford DJ**, **Rusanov D**, **Winestone B**, **Kaplan R**, **Fuller-Tyszkiewicz M**, **Melvin G** (2023). Disclosure of suicidal ideation and behaviours: A systematic review and meta-analysis of prevalence. *Clinical Psychology Review* **101**, 102272.

[60] **Hay N**, **Russell SJ** (2011). Metareasoning for monte carlo tree search. Technical report, Electrical Engineering and Computer Sciences, U of C at Berkeley.

[61] **Hewitt SR**, **Norbury A**, **Huys QJ**, **Hauser TU** (2025). Real-world fluctuations in motivation drive effort-based choices. *Proc Natl Acad Sci U S A* **122**(12), e2417964122.

[62] **Hjort MM**, **Garrett ZQ**, **Gordon AG**, **Ancell E**, **Trzeciak M**, **Lu PY**, **Bruchas MR**, **Witten DM**, **Steinmetz NA**, **Stuber GD** (2026). Prefrontal to ventral tegmental area dynamics drive contingency degradation. *Nature* pages 1–9.

[63] **Hoemann K**, **Xu F**, **Barrett LF** (2019). Emotion words, emotion concepts, and emotional development in children: A constructionist hypothesis. *Developmental Psychology* **55**(9), 1830–1849.

[64] **Hommel BE**, **Arslan RC** (2025). Language models accurately infer correlations between psychological items and scales from text alone. *Advances in Methods and Practices in Psychological Science* **8**(4).

[65] **Hua Y**, **Siddals S**, **Ma Z**, **Galatzer-Levy I**, **Xia W**, **Hau C**, **Na H**, **Flathers M**, **Linardon J**, **Ayubcha C**, **Torous J** (2025). Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: a systematic review. *World Psychiatry* **24**(3), 383–394.

[66] **Huang S**, **Pollak SD**, **Xie W** (2025). Conceptual knowledge increasingly supports emotion understanding as perceptual contribution declines with age. *Nature Communications* **16**(1).

[67] **Hur JK**, **Heffner J**, **Feng GW**, **Joormann J**, **Rutledge RB** (2024). Language sentiment predicts changes in depressive symptoms. *Proceedings of the National Academy of Sciences* **121**(39), e2321321121.

[68] **Huth AG**, **de Heer WA**, **Griffiths TL**, **Theunissen FE**, **Gallant JL** (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**(7600), 453–458.

[69] **Huys QJM**, **Dayan P**, **Daw** (2015). Depression: A Decision-Theoretic Account. *Ann. Rev. Neurosci.* **38**, 1–23.

[70] **Huys QJM**, **Eshel N**, **O'Nions E**, **Sheridan L**, **Dayan P**, **Roiser JP** (2012). Bonsai trees in your head: how the Pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Comput Biol* **8**(3), e1002410.

[71] **Huys QJM**, **Renz D** (2017). A formal valuation framework for emotions and their control. *Biological Psychiatry* **82**, 413–420.

[72] **Ibrahim L**, **Hafner FS**, **Rocher L** (2026). Training language models to be warm can reduce accuracy and increase sycophancy. *Nature* **652**(8112), 1159–1165.

[73] **Irie K**, **Csordás R**, **Schmidhuber J** (2022). The dual form of neural networks revisited: Connecting test time predictions to training patterns via spotlights of attention. In K Chaudhuri, S Jegelka, L Song, C Szepesvari, G Niu, S Sabato, eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9639–9659. PMLR.

[74] **Iwata J**, **LeDoux JE** (1988). Dissociation of associative and nonassociative concomitants of classical fear conditioning in the freely behaving rat. *Behavioral neuroscience* **102**, 66–76.

[75] **Jagadish AK**, **Thalmann M**, **Coda-Forno J**, **Binz M**, **Schulz E** (2025). Meta-learning ecological priors from large language models explains human learning and decision making. ArXiv:2509.00116 [q-bio].

[76] **James W** (1884). *What is emotion?*, pages 290–303. Appleton-Century-Crofts.

[77] **Jiang Y**, **Rajendran G**, **Ravikumar AP**, **Aragam B**, **Veitch V** (2024). On the origins of linear representations in large language models.

[78] **Kim J**, **Ma SP**, **Chen ML**, **Galatzer-Levy IR**, **Torous J**, **van Roessel PJ**, **Sharp C**, **Pfeffer MA**, **Rodriguez CI**, **Linos E**, **Chen JH** (2025). Optimizing large language models for detecting symptoms of depression/anxiety in chronic diseases patient communications. *npj Digital Medicine* **8**(1), 580.

[79] **Kirsch L**, **Harrison J**, **Sohl-Dickstein J**, **Metz L** (2024). General-Purpose In-Context Learning by Meta-Learning Transformers. ArXiv:2212.04458 [cs].

[80] **Kjell ONE**, **Sikström S**, **Kjell K**, **Schwartz HA** (2022). Natural language analyzed with ai-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. *Scientific Reports* **12**(1).

[81] **Kragel PA**, **LaBar KS** (2016). Decoding the Nature of Emotion in the Brain. *Trends in Cognitive Sciences* **20**(6), 444–455.

[82] **Kroenke K**, **Spitzer RL**, **Williams JB** (2001). The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine* **16**, 606–613.

[83] **Lampinen AK**, **Chan SCY**, **Singh AK**, **Shanahan M** (2025). The broader spectrum of in-context learning. ArXiv:2412.03782 [cs].

[84] **Lazarus RS** (1991). Progress on a cognitive-motivational-relational theory of emotion. *American Psychologist* **46**(8), 819–834.

[85] **Lerner JS**, **Keltner D** (2000). Beyond valence: Toward a model of emotion-specific influences on judgement and choice. *Cognition & Emotion* **14**(4), 473–493.

[86] **Lerner JS**, **Li Y**, **Valdesolo P**, **Kassam KS** (2015). Emotion and decision making. *Annual review of psychology* **66**, 799–823.

[87] **Li HX**, **Lu B**, **Wang YW**, **Li XY**, **Chen X**, **Yan CG** (2023). Neural representations of self-generated thought during think-aloud fMRI. *NeuroImage* **265**, 119775.

[88] **Li M**, **Su Y**, **Huang HY**, **Cheng J**, **Hu X**, **Zhang X**, **Wang H**, **Qin Y**, **Wang X**, **Lindquist KA**, **Liu Z**, **Zhang D** (2024). Language-specific representation of emotion-concept knowledge causally supports emotion inference. *iScience* **27**(12).

[89] **Li X**, **Zhang Y**, **Tiwari P**, **Song D**, **Hu B**, **Yang M**, **Zhao Z**, **Kumar N**, **Marttinen P** (2022). EEG Based Emotion Recognition: A Tutorial and Review. *ACM Comput. Surv.* **55**(4), 79:1–79:57.

[90] **Li Y**, **Chen J**, **Li F**, **Fu B**, **Wu H**, **Ji Y**, **Zhou Y**, **Niu Y**, **Shi G**, **Zheng W** (2023). GMSS: Graph-Based Multi-Task Self-Supervised Learning for EEG Emotion Recognition. *IEEE Transactions on Affective Computing* **14**(3), 2512–2525.

[91] **Lieder F**, **Callaway F**, **Griffiths T** (2025). The rational use of cognitive resources .

[92] **Lim SM**, **Shiau CWC**, **Cheng LJ**, **Lau Y** (2022). Chatbot-Delivered Psychotherapy for Adults With Depressive and Anxiety Symptoms: A Systematic Review and Meta-Regression. *Behavior Therapy* **53**(2), 334–347.

[93] **Lindquist KA**, **Wager TD**, **Kober H**, **Bliss-Moreau E**, **Barrett LF** (2012). The brain basis of emotion: A meta-analytic review. *The Behavioral and brain sciences* **35**(3), 121–143.

[94] **Linehan MM** (1993). *Skills training manual for treating borderline personality disorder.* Guilford Press.

[95] **Loewenstein G** (1996). Out of control: Visceral influences on behavior. *Organizational behavior and human decision processes* **65**(3), 272–292.

[96] **Low DM**, **Rumker L**, **Talkar T**, **Torous J**, **Cecchi G**, **Ghosh SS** (2020). Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study. *Journal of Medical Internet Research* **22**(10), e22635.

[97] **Lupyan G**, **Abdel Rahman R**, **Boroditsky L**, **Clark A** (2020). Effects of language on visual perception. *Trends in Cognitive Sciences* **24**(11), 930–944.

[98] **Mahowald K**, **Ivanova AA**, **Blank IA**, **Kanwisher N**, **Tenenbaum JB**, **Fedorenko E** (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences* **0**(0).

[99] **Mao H**, **Liu G**, **Ma Y**, **Wang R**, **Johnson K**, **Tang J** (2025). A survey to recent progress towards understanding in-context learning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7302–7323. Association for Computational Linguistics.

[100] **Marr D** (1982). *Vision*. Freeman, New York, NY, USA.

[101] **Mathews A**, **MacLeod C** (1985). Selective processing of threat cues in anxiety states. *Behav Res Ther* **23**(5), 563–569.

[102] **Mathews A**, **Ridgeway V**, **Williamson DA** (1996). Evidence for attention to threatening stimuli in depression. *Behav Res Ther* **34**(9), 695–705.

[103] **Mathys C**, **Daunizeau J**, **Friston KJ**, **Stephan KE** (2011). A bayesian foundation for individual learning under uncertainty. *Frontiers in human neuroscience* **5**, 39.

[104] **Mattar MG**, **Daw ND** (2018). Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience* **21**, 1609–1617.

[105] **Miranda B**, **Butler JL**, **Malalasekera WMN**, **Behrens TE**, **Dayan P**, **Kennerley SW** (2026). Neural signatures of model-based and model-free reinforcement learning across prefrontal cortex and striatum. *eLife* **14**.

[106] **Moon S**, **Lee A**, **Kim JE**, **Kang HJ**, **Shin IS**, **Kim SW**, **Kim JM**, **Jhon M**, **Kim JW** (2025). DepressLLM: Interpretable domain-adapted language model for depression detection from real-world narratives. ArXiv:2508.08591 [cs].

[107] **Moors A**, **Ellsworth PC**, **Scherer KR**, **Frijda NH** (2013). Appraisal theories of emotion: State of the art and future development. *Emotion Review* **5**(2), 119–124.

[108] **Moutoussis M**, **Bentall RP**, **Williams J**, **Dayan P** (2008). A temporal difference account of avoidance learning. *Network* **19**(2), 137–160.

[109] **Nepal S**, **Pillai A**, **Campbell W**, **Massachi T**, **Heinz MV**, **Kunwar A**, **Choi ES**, **Xu X**, **Kuc J**, **Huckins JF**, **Holden J**, **Preum SM**, **Depp C**, **Jacobson N**, **Czerwinski MP**, **Granholm E**, **Campbell AT** (2024). MindScape Study: Integrating LLM and Behavioral Sensing for Personalized AI-Driven Journaling Experiences. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **8**(4), 186:1–186:44.

[110] **Nie J**, **Shao HV**, **Fan Y**, **Shao Q**, **You H**, **Preindl M**, **Jiang X** (2025). LLM-based Conversational AI Therapist for Daily Functioning Screening and Psychotherapeutic Intervention via Everyday Smart Devices. *ACM Transactions on Computing for Healthcare* page 3712299.

[111] **Niedenthal PM**, **Showers C** (1991). The Perception and Processing of Affective Information and its Influences on Social Judgment. In *Emotion and Social Judgements*. Garland Science.

[112] **Norbury A**, **Dercon Q**, **Hauser TU**, **Dolan RJ**, **Huys QJM** (2025). Learning training as a cognitive restructuring intervention. *Biological psychiatry: Cognitive neuroscience and neuroimaging.* page Online ahead of print.

[113] **Norbury A**, **Hauser T**, **Fleming S**, **Dolan R**, **Huys QJM** (2024). Different components of cognitive-behavioural therapy affect specific cognitive mechanisms. *Science Advances* **10**, eadk3222.

[114] **Olsson C**, **Elhage N**, **Nanda N**, **Joseph N**, **DasSarma N**, **Henighan T**, **Mann B**, **Askell A**, **Bai Y**, **Chen A**, **Conerly T**, **Drain D**, **Ganguli D**, **Hatfield-Dodds Z**, **Hernandez D**, **Johnston S**, **Jones A**, **Kernion J**, **Lovitt L**, **Ndousse K**, **Amodei D**, **Brown T**, **Clark J**, **Kaplan J**, **McCandlish S**, **Olah C** (2022). In-context learning and induction heads.

[115] **Onysk J**, **Huys QJM** (2025). Quantifying depressive mental states with large language models. ArXiv:2502.09487 [cs].

[116] **Ortega PA**, **Wang JX**, **Rowland M**, **Genewein T**, **Kurth-Nelson Z**, **Pascanu R**, **Heess N**, **Veness J**, **Pritzel A**, **Sprechmann P**, **Jayakumar SM**, **McGrath T**, **Miller K**, **Azar M**, **Osband I**, **Rabinowitz N**, **György A**, **Chiappa S**, **Osindero S**, **Teh YW**, **Hasselt Hv**, **Freitas Nd**, **Botvinick M**, **Legg S** (2019). Meta-learning of Sequential Strategies. ArXiv:1905.03030 [cs].

[117] **Palminteri S**, **Wyart V**, **Koechlin E** (2017). The importance of falsification in computational cognitive modeling. *Trends in cognitive sciences* **21**, 425–433.

[118] **Panksepp J** (1998). *Affective Neuroscience*. OUP, New York, NY.

[119] **Panksepp J** (2007). Neurologizing the psychology of affects: How appraisal-based constructivism and basic emotion theory can coexist. *Perspectives on psychological science : a journal of the Association for Psychological Science* **2**, 281–296.

[120] **Pessoa L** (2017). A Network Model of the Emotional Brain. *Trends in Cognitive Sciences* **21**(5), 357–371.

[121] **Phuong M**, **Hutter M** (2022). Formal algorithms for transformers ArXiv:2207.09238 [cs].

[122] **Pittig A**, **Treanor M**, **LeBeau RT**, **Craske MG** (2018). The role of associative fear and avoidance learning in anxiety disorders: Gaps and directions for future research. *Neuroscience & Biobehavioral Reviews* **88**, 117–140.

[123] **Reichman B**, **Avsian A**, **Webster S**, **Heck L** (2026). Emotion is not just a label: Latent emotional factors in llm processing.

[124] **Ren J**, **Guo Q**, **Yan H**, **Liu D**, **Zhang Q**, **Qiu X**, **Lin D** (2024). Identifying semantic induction heads to understand in-context learning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6916–6932. Association for Computational Linguistics.

[125] **Ren Z**, **Yang Z**, **Ye C**, **Sun H**, **Chen C**, **Zhu X**, **Liao X** (2025). Fine-grained emotion recognition via in-context learning. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, pages 2503–2513. ACM.

[126] **Russell JA** (1980). A circumplex model of affect. *Journal of Personality and Social Psychology* **39**(6), 1161–1178.

[127] **Russell S**, **Wefald E** (1991). Principles of metareasoning. *Artificial intelligence* **49**(1-3), 361–395.

[128] **Russell SJ**, **Wefald EH** (1991). *Do the right thing: studies in limited rationality*. MIT press.

[129] **Rutledge RB**, **Skandali N**, **Dayan P**, **Dolan RJ** (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 12252–12257.

[130] **Satpute AB**, **Shu J**, **Weber J**, **Roy M**, **Ochsner KN** (2013). The functional neural architecture of self-reports of affective experience. *Biological Psychiatry* **73**(7), 631–638.

[131] **Scherer KR** (1984). Emotion as a multicomponent process: A model and some cross-cultural data. *Review of Personality & Social Psychology* **5**, 37–63.

[132] **Scherer KR** (2005). What are emotions? and how can they be measured? *Social science information* **44**(4), 695–729.

[133] **Schlagenhauf F**, **Huys QJM**, **Deserno L**, **Rapp MA**, **Beck A**, **Heinze HJ**, **Dolan R**, **Heinz A** (2014). Striatal dysfunction during reversal learning in unmedicated schizophrenia patients. *Neuroimage* **89**, 171–180.

[134] **Scholich T**, **Barr M**, **Stirman SW**, **Raj S** (2025). A Comparison of Responses from Human Therapists and Large Language Model–Based Chatbots to Assess Therapeutic Communication: Mixed Methods Study. *JMIR Mental Health* **12**(1), e69709.

[135] **Schupp HT**, **Flaisch T**, **Stockburger J**, **Junghöfer M** (2006). Emotion and attention: event-related brain potential studies. *Progress in Brain Research* **156**, 31–51.

[136] **Seabrook EM**, **Kern ML**, **Fulcher BD**, **Rickard NS** (2018). Predicting Depression From Language-Based Emotion Dynamics: Longitudinal Analysis of Facebook and Twitter Status Updates. *Journal of Medical Internet Research* **20**(5), e9267.

[137] **Seymour B**, **Dolan R** (2008). Emotion, Decision Making, and the Amygdala. *Neuron* **58**(5), 662–671.

[138] **Shen T**, **Dayan P** (2025). Individual differences in tail risk sensitive exploration using bayes-adaptive markov decision processes. *eLife* **13**.

[139] **Shin D**, **Kim H**, **Lee S**, **Cho Y**, **Jung W** (2024). Using Large Language Models to Detect Depression From User-Generated Diary Text Data as a Novel Approach in Digital Mental Health Screening: Instrument Validation Study. *Journal of Medical Internet Research* **26**(1), e54617.

[140] **Siegel EH**, **Sands MK**, **Van den Noortgate W**, **Condon P**, **Chang Y**, **Dy J**, **Quigley KS**, **Barrett LF** (2018). Emotion fingerprints or emotion populations? A meta-analytic investigation of autonomic features of emotion categories. *Psychological Bulletin* **144**(4), 343–393.

[141] **Siegel EH**, **Wormwood JB**, **Quigley KS**, **Barrett LF** (2018). Seeing What You Feel: Affect Drives Visual Perception of Structurally Neutral Faces. *Psychological Science* **29**(4), 496–503.

[142] **Simon HA** (1955). A behavioral model of rational choice. *The quarterly journal of economics* pages 99–118.

[143] **Sosea T**, **Caragea C** (2025). Hard Emotion Test Evaluation Sets for Language Models. In L Chiruzzo, A Ritter, L Wang, eds., *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7930–7944. Association for Computational Linguistics, Albuquerque, New Mexico.

[144] **Stade EC**, **Stirman SW**, **Ungar LH**, **Boland CL**, **Schwartz HA**, **Yaden DB**, **Sedoc Ja**, **DeRubeis RJ**, **Willer R**, **Eichstaedt JC** (2024). Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *npj Mental Health Research* **3**(1), 12.

[145] **Stanley J**, **Rabot E**, **Reddy S**, **Belilovsky E**, **Mottron L**, **Bzdok D** (2025). Large language models deconstruct the clinical intuition behind diagnosing autism. *Cell* **188**(8), 2235–2248.e10.

[146] **Sun X**, **Comrie AE**, **Kahn AE**, **Monroe EJ**, **Washington CB**, **Joshi A**, **Guidera JA**, **Denovellis EL**, **Krausz TA**, **Zhou J**, **Thompson P**, **Hernandez J**, **Yorita A**, **Haque R**, **Pandarinath C**, **Berke JD**, **Daw ND**, **Frank LM** (2026). Meta-learning is expressed through altered prefrontal cortical dynamics.

[147] **Tang J**, **LeBel A**, **Jain S**, **Huth AG** (2023). Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience* **26**(5), 858–866.

[148] **Tikochinski R**, **Goldstein A**, **Meiri Y**, **Hasson U**, **Reichart R** (2025). Incremental accumulation of linguistic context in artificial and biological neural networks. *Nature Communications* **16**(1), 803.

[149] **Tuckute G**, **Kanwisher N**, **Fedorenko E** (2024). Language in brains, minds, and machines. *Annual Review of Neuroscience* **47**(Volume 47, 2024), 277–301.

[150] **Tuckute G**, **Sathe A**, **Srikant S**, **Taliaferro M**, **Wang M**, **Schrimpf M**, **Kay K**, **Fedorenko E** (2024). Driving and suppressing the human language network using large language models. *Nature Human Behaviour* **8**(3), 544–561.

[151] **Turner RE** (2023). An introduction to transformers. ArXiv:2304.10557 [cs].

[152] **Tversky A**, **Kahneman D** (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science* **185**(4157), 1124–1131.

[153] **van Duijn M**, **van Dijk B**, **Kouwenhoven T**, **de Valk W**, **Spruit M**, **van der Putten P** (2023). Theory of Mind in Large Language Models: Examining Performance of 11 State-of-the-Art models vs. Children Aged 7-10 on Advanced Tests. In J Jiang, D Reitter, S Deng, eds., *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 389–402. Association for Computational Linguistics, Singapore.

[154] **van Opheusden B**, **Kuperwajs I**, **Galbiati G**, **Bnaya Z**, **Li Y**, **Ma WJ** (2023). Expertise increases planning depth in human gameplay. *Nature* **618**, 1000–1005.

[155] **Vaswani A**, **Shazeer N**, **Parmar N**, **Uszkoreit J**, **Jones L**, **Gomez AN**, **Kaiser L**, **Polosukhin I** (2017). Attention Is All You Need. ArXiv:1706.03762 [cs] version: 5.

[156] **Wang JX**, **Kurth-Nelson Z**, **Kumaran D**, **Tirumala D**, **Soyer H**, **Leibo JZ**, **Hassabis D**, **Botvinick M** (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience* **21**(6), 860–868.

[157] **Wang Z**, **Wang Y**, **Hu C**, **Yin Z**, **Song Y** (2022). Transformers for EEG-Based Emotion Recognition: A Hierarchical Spatial Information Learning Model. *IEEE Sensors Journal* **22**(5), 4359–4368.

[158] **Watson D**, **Clark LA**, **Tellegen A** (1988). Development and validation of brief measures of positive and negative affect: The panas scales. *Journal of Personality and Social Psychology* **54**(6), 1063–1070.

[159] **Weber I**, **Zorowitz S**, **Niv Y**, **Bennett D** (2022). The effects of induced positive and negative affect on pavlovian-instrumental interactions. *Cognition & Emotion* pages 1–18.

[160] **Wells A**, **Fisher P**, **Myers S**, **Wheatley J**, **Patel T**, **Brewin CR** (2009). Metacognitive therapy in recurrent and persistent depression: A multiple-baseline study of a new treatment. *Cognitive Therapy and Research* **33**(3), 291–300.

[161] **Wright AGC**, **Ringwald WR**, **Vize CE**, **Eichstaedt JC**, **Angstadt M**, **Taxali A**, **Sripada C** (2026). Assessing personality using zero-shot generative ai scoring of brief open-ended text. *Nature Human Behaviour* **10**(3), 541–555.

[162] **Wulff DU**, **Mata R** (2025). Semantic embeddings reveal and address taxonomic incommensurability in psychological measurement. *Nature Human Behaviour* **9**(5), 944–954.

[163] **Wurgaft D**, **Lubana ES**, **Park CF**, **Tanaka H**, **Reddy G**, **Goodman ND** (2025). In-Context Learning Strategies Emerge Rationally. ArXiv:2506.17859 [cs].

[164] **Xie SM**, **Min S** (2022). How does in-context learning work? a framework for understanding the differences from traditional supervised learning. The Stanford AI Lab Blog.

[165] **Xie SM**, **Raghunathan A**, **Liang P**, **Ma T** (2021). An explanation of in-context learning as implicit bayesian inference. ArXiv:2111.02080 [cs].

[166] **Xu X**, **Yao B**, **Dong Y**, **Gabriel S**, **Yu H**, **Hendler J**, **Ghassemi M**, **Dey AK**, **Wang D** (2024). Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **8**(1), 1–32.

[167] **Zadra JR**, **Clore GL** (2011). Emotion and perception: the role of affective information. *Wiley interdisciplinary reviews. Cognitive science* **2**, 676–685.

[168] **Zani P**, **Jones T**, **Neuhaus R**, **Milgrom J** (2009). Effect of refuge distance on escape behavior of side-blotched lizards (uta stansburiana). *Canadian Journal of Zoology* **87**(5), 407–414.

[169] **Zhang J**, **Zhong L** (2025). Decoding Emotion in the Deep: A Systematic Study of How LLMs Represent, Retain, and Express Emotion. ArXiv:2510.04064 [cs].

[170] **Zhang Y**, **Gibson E**, **Davis F** (2023). Can Language Models Be Tricked by Language Illusions? Easier with Syntax, Harder with Semantics. In J Jiang, D Reitter, S Deng, eds., *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 1–14. Association for Computational Linguistics, Singapore.

[171] **Zhang Y**, **Zhang F**, **Yang Z**, **Wang Z** (2023). What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization.

[172] **Zhou F**, **Zhao W**, **Qi Z**, **Geng Y**, **Yao S**, **Kendrick KM**, **Wager TD**, **Becker B** (2021). A distributed fMRI-based signature for the subjective experience of fear. *Nature Communications* **12**(1), 6643.

[173] **Zhu J**, **Ji L**, **Liu C** (2019). Heart rate variability monitoring for emotion and disorders of emotion. *Physiological Measurement* **40**(6), 064004.