How local and global metacognition shape mental health

Tricia X.F. Seow, Marion Rouault, Claire M. Gillan, Stephen M. Fleming

PII: S0006-3223(21)01329-9
DOI: https://doi.org/10.1016/j.biopsych.2021.05.013
Reference: BPS 14562

To appear in: Biological Psychiatry

Received Date: 18 November 2020
Revised Date: 14 May 2021
Accepted Date: 16 May 2021

Please cite this article as: Seow T.X.F., Rouault M., Gillan C.M. & Fleming S.M., How local and global metacognition shape mental health, Biological Psychiatry (2021), doi: https://doi.org/10.1016/j.biopsych.2021.05.013.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2021 Published by Elsevier Inc on behalf of Society of Biological Psychiatry.
TITLE
How local and global metacognition shape mental health

SHORT TITLE
How local and global metacognition shape mental health

AUTHOR LIST
Tricia X.F. Seow\textsuperscript{1,2,*}, Marion Rouault\textsuperscript{3,4,*}, Claire M. Gillan\textsuperscript{5} and Stephen M. Fleming\textsuperscript{1,2,6} (* indicates equal contribution)

AUTHOR AFFILIATIONS
\textsuperscript{1}Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, 10-12 Russell Square, London WC1B 5EH, UK.
\textsuperscript{2}Wellcome Centre for Human Neuroimaging, University College London, 12 Queen Square, London WC1N 3AR, UK.
\textsuperscript{3}Institut Jean Nicod, Département d'études cognitives, ENS, EHESS, CNRS, PSL University, 75005 Paris, France.
\textsuperscript{4}Laboratoire de neurosciences cognitives et computationnelles, Département d'études cognitives, ENS, INSERM, PSL University, 75005 Paris, France.
\textsuperscript{5}School of Psychology, Trinity College Dublin, Dublin, Ireland.
\textsuperscript{6}Department of Experimental Psychology, University College London, 26 Bedford Way, London WC1H 0AP, UK.

NAME AND CONTACT INFORMATION OF CORRESPONDING AUTHORS
Marion Rouault and Tricia X.F. Seow

- Marion Rouault: Département d'Études Cognitives, École Normale Supérieure, 29 rue d'Ulm, 75005 Paris, France. Email: marion.rouault@gmail.com.
- Tricia X.F. Seow: Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, 10-12 Russell Square, WC1B 5EH London, UK. Email: t.seow@ucl.ac.uk

KEYWORDS
Metacognition, self-beliefs, self-efficacy, confidence, mental health, transdiagnostic psychiatry
**ABSTRACT**

Metacognition is the ability to reflect on our own cognition and mental states. It is a critical aspect of human subjective experience and operates across many hierarchical levels of abstraction—encompassing “local” confidence in isolated decisions and “global” self-beliefs about our abilities and skills. Alterations in metacognition are considered foundational to neurological and psychiatric disorders, but research has mostly focused on local metacognitive computations, missing out on the role of global aspects of metacognition. Here, we first review current behavioral and neural metrics of local metacognition that lay the foundation for this research. We then address the neurocognitive underpinnings of global metacognition uncovered by recent studies. Finally, we outline a theoretical framework in which higher hierarchical levels of metacognition may help identify the role of maladaptive metacognitive evaluation in mental health conditions, particularly when combined with transdiagnostic methods.
INTRODUCTION

Metacognition, the ability to reflect on and evaluate our own thoughts and actions, is a crucial component of human behavior and subjective experience (1). A wealth of empirical studies have shown that impaired metacognition is associated with detrimental behavior and poor mental health (2,3). For instance, delusional thinking in schizophrenic patients is thought to be maintained by metacognitive deficits such as a lack of insight (4) or overconfidence in incorrect models of the world (5–7). In a range of mental health conditions, metacognition shows consistent, yet specific, individual differences (8,9) (see review (2)), findings that generalize across various tasks (10) and cognitive domains (11), and abnormalities that may be heritable (12). As researchers in psychiatry aim to develop reliable neurocognitive markers for identifying current and future mental health problems, metacognitive assessments hold promise (13).

There are several challenges in meeting this aim. First, metacognition is tightly bound to cognitive performance, such as the accuracy of visual decisions or memory recollection. Second, metacognition manifests in various hierarchical levels of abstraction, from “local” confidence in isolated decisions to more “global” metacognitive constructs like self-efficacy beliefs. Whilst most research has focused on local metacognition, we propose that global aspects of metacognition may be more closely related to daily functioning and the subjective experience of mental health symptoms. Lastly, metacognitive changes may not be readily apparent in case-control comparisons using standard diagnostic categories and instead be better captured by transdiagnostic dimensions. Here, we introduce the main behavioral and neural metrics of local metacognition, discuss the relevance of global metacognition for mental health and outline how transdiagnostic methodologies may help to unpack the role of multiple hierarchical levels of metacognition in psychiatry. Note that the disorders which we raise as examples are those with the greatest relevance to the transdiagnostic studies we discuss later.

Section 1: Methods for quantifying local metacognition

Behavioral and computational metrics of local metacognition

Several metrics have been developed to quantify local metacognition in laboratory tasks, most of which rely on examining the correspondence between objective performance and confidence ratings (a subjective report of being correct about a decision/statement...
across multiple experimental trials (Figure 1A). Two independent aspects of local metacognition can be distinguished: metacognitive bias and sensitivity (15) (Figure 1B). Metacognitive bias reflects how confident we are irrespective of actual performance, and is usually estimated as the mean confidence rating averaged over correct and incorrect judgements. In contrast, metacognitive sensitivity reflects an ability to discriminate correct from incorrect judgements. A participant who rates high confidence on correct judgements and low confidence on incorrect judgements is estimated to have high metacognitive sensitivity.

An initial wave of studies relied on simple correlation statistics, which conflated metacognitive bias and sensitivity in one measure, an issue covered previously (16,17). More recent methods (i.e., type 2 signal detection theory (SDT)) estimate a bias-free assessment of metacognitive sensitivity (18). However, metacognitive sensitivity is typically dependent on task performance, where easier tasks produce greater sensitivity (16). Instead, model-based methods reliant on SDT (e.g., meta-d’ model) correct for such performance confounds, leading to the derivation of summary statistics such as metacognitive efficiency that represent a participant’s level of metacognitive sensitivity corrected for variation in task performance (17). Another approach is to use staircase procedures (19,20) that adjust task performance at a predetermined level, allowing variation in metacognitive sensitivity to be isolated (e.g., (8)), although this method has caveats (21). Failures to replicate metamemory biases towards lowered confidence in obsessive-compulsive disorder (OCD) (22–31) or recent evidence of previously inflated effects (32) of higher confidence in errors from delusion-prone and paranoid schizophrenic patients (5,6,33–38) were ultimately explained by metacognitive sensitivity and bias not being properly separated. Future experiments should aim to minimize potential confounds in estimating metacognitive sensitivity at either the paradigm design or analysis stage.

Neural bases of local metacognition
Beyond behavioral metrics, studies have begun to reveal the neural bases of local metacognition about perception and memory (see review (11)). Strong convergent evidence highlights the importance of prefrontal cortex (PFC) for metacognition. Lesions (39) or transcranial magnetic stimulation (40) to the PFC affect perceptual metacognitive sensitivity while leaving task performance unaffected. Structural and functional MRI
studies in healthy humans have linked individual differences in anterior PFC volume, function and connectivity to metacognitive ability (8,20,41–47). Beyond PFC, a distributed network of brain regions including the medial PFC, precuneus and hippocampus (20,43,44,46–52) are also involved in metacognition. Electrophysiology studies provided convergent evidence of activity associated with metacognition in prefrontal theta oscillations (53), the P3 ERP component (54) and the error-related negativity (ERN) (55–57). Similar neural correlates are observed in relation to aberrant metacognitive processes in some psychiatric disorders. Altered metacognition about perceptual decisions in schizophrenia patients correlates with hypoactivity in fronto-parietal areas (58), and also hippocampal volume and its grey matter microstructure (59). Drug addiction, which was linked to deficits in error awareness (60) and perceptual metacognitive sensitivity (61), was linked to hypoactivity and a loss of structural integrity in the anterior cingulate cortex. Overall, the medial PFC and parietal cortex are proposed to play a domain-general role in metacognition, with other nodes of the network contributing in a domain-specific fashion (11) (Figure 2).

Section 2: From local confidence to global self-beliefs

Many forms of metacognition co-exist

While the psychological and neural bases of local metacognition are increasingly well characterized, its functional roles remain less clear. Local confidence has been suggested to regulate subsequent decisions by recruiting cognitive control (62), gathering information (63), controlling exploration (64) and adapting speed-accuracy trade-offs (65). However, these are all limited in scope and on short time scales. In contrast to local confidence in single decisions, global metacognitive evaluations of performance (“self-beliefs”) can span several decisions or experimental trials, allowing for a gradual formation of self-performance estimates in numerous aspects: about our ability on a given task, in a specific cognitive domain, or even how capable we feel, broadly (Figure 3). In turn, these self-beliefs may affect future decisions on longer time scales (66,67), such as promoting the initiation of behavioral sequences towards achieving a goal. Individuals with low self-beliefs tend to feel less in control of their environment, are less likely to believe that their decisions will affect future outcomes, and are slower to recover after setbacks (68,69). Accordingly, distorted self-beliefs may have a pervasive impact in educational and clinical settings (70), determining how people see themselves and their capabilities. But despite their recognized importance
for mental health, the cognitive and neural foundations of self-beliefs remain largely unclear.

Self-beliefs are related to the psychological construct of self-esteem, a global notion of self-worth that cuts across many domains (e.g. physical, social and academic) (71). Low self-esteem is a key predictor of mental health issues such as anxiety and depression (72,73). Low self-esteem has strong theoretical ties to dominant clinical psychology models of depression (74), where depressive symptoms are thought to be grounded in negative schema that persist despite alternative evidence (75). Negative schemas encompass several processes, among which confidence/self-beliefs are one critical aspect, with the proposed neural correlates of negative schemas and confidence partly overlapping, e.g. cingulate cortex (76). However, despite the strong face validity of these negative schema, their measurement with clinical scales precludes a mechanistic understanding of how these self-reports arise (77). In contrast, models of global metacognition constitute a mechanistic framework within which to define testable hypotheses, and unpack the mechanisms underpinning low self-beliefs. For instance, we can examine how shifts in processes supporting local decision confidence lead to gradual changes in global self-beliefs that likely unfold over longer timescales. The study of apathy provides a recent example—a single self-report (i.e., apathetic state) could be attributed to various computational mechanisms (reduced reward sensitivity or increased subjective perception of effort), each associated with distinct neurobiological systems (78,79).

**Neurocognitive foundations of simple forms of global self-beliefs**

We have begun to delineate computations underlying the formation of global self-beliefs from local confidence estimates (80,81). In these experiments, participants were asked to perform mini blocks of two interleaved perceptual tasks. At the end of each block, they selected the task which they thought they performed best—a proxy for global self-beliefs about the two tasks. Local subjective confidence ratings were found to predict global self-beliefs over and above objective performance (80). Using functional MRI, we further found that ventral striatal activity reflected the level of global self-beliefs (but not local confidence signals), while confidence-related activity in ventromedial PFC (vmPFC) was further modulated by the level of global self-belief (81). This is in line with two studies indicating that vmPFC reflects fluctuations in self-performance estimates on mini games.
performed across several trials when participants monitor expected task success with (64) or without (82) external feedback. Moreover, white matter structural integrity between ventral striatum and vmPFC, estimated using DTI, shows systematic links with individual self-esteem (83). These results establish an initial link between local and global metacognition (Figure 2), and reveal neural representations of global self-beliefs that go beyond the tracking of local confidence (84).

It is important to acknowledge that global self-beliefs assessed in these studies were limited to the scope of a lab experiment, and to perceptual (80,84) or color/time estimation tasks (82). These tasks are well characterized in terms of how local perceptual decisions and confidence estimates are formed (e.g., (85)), which is vital for precisely quantifying how self-beliefs are constructed from local confidence and external feedback (88). However, there is a substantial gap between experimental investigations of so-called global self-beliefs and self-beliefs relevant to real-life decisions, which typically fluctuate over considerably longer time scales than those assessed in the laboratory. Additionally, many other factors contribute to the formation of real-life self-beliefs, such as feedback from other people and one’s social environment (86,87). We suggest that we can bridge the gap by examining how self-beliefs generalize across different tasks and across cognitive domains (Figure 3). Such a generalization mechanism should normally support the formation of useful priors about expected ability in closely related tasks. But if this mechanism becomes maladaptive, leading to, e.g., excessive generalization from local experiences, it could create volatile self-beliefs. Conversely, a disruption in updating mechanisms could result in rigid self-beliefs being insufficiently updated in light of new positive experiences.

Relating global self-beliefs to functional symptoms
Adapting a framework for global metacognition may prove useful clinically because it may be more directly relevant to the subjective and functional experiences of patients as compared to local confidence in isolated decisions. For example, anosognosia, defined as a lack of awareness of cognitive deficits, particularly about memory, is a common symptom of dementia (88). A lack of self-awareness may lead to a failure to adapt to changes in cognitive abilities, for instance leading to risky behaviors such as driving long distances or traveling to unfamiliar locations (89). Anosognosia may also affect decisions about appropriate courses of treatment or prevent the implementation of
strategies to aid memory such as setting reminders (90,91). Similarly, intact global metacognition may be crucial for treatment adherence as an individual may only be willing to participate in therapeutic interventions if they have insight into their symptoms. Previous work with schizophrenic patients has indeed shown that clinical insight is predictive of medication compliance (92,93).

At present, only local confidence is routinely measured in experimental studies of metacognition. However, there is likely a complex and largely unexplored interplay between local metacognitive evaluations and global self-beliefs. Notably, anosognosia may co-exist with relatively intact local metacognition about performance on individual trials. In these studies (94–96), participants with Alzheimer’s disease underwent assessments of local metacognition on memory and motor tasks, and clinicians evaluated the patients’ global awareness of their deficits (95). While both local memory and motor metacognition were found to be relatively intact (89,95), there was a specific deficit of global awareness in the memory (and not motor) domain (95), suggesting that local and global metacognitive levels may dissociate in some cognitive domains, but not others. We note, however, that extended clinical interviews and/or informants’ reports were used as proxies for ground-truth ability; as such, the data remains disconnected from approaches that seek to model the relationship between performance and confidence.

Global and local metacognition also diverge in Parkinson’s disease. Patients differ from healthy participants in their feeling of knowing accuracy in recognition memory tests at the item level, but not in their global prediction of accuracy (97). These examples highlight the value of a neurocognitive framework encompassing local and global metacognition, to pinpoint the origins of lack of awareness (80). It could be that symptom severity only affects upper hierarchical levels (Figure 3), or creates imbalances between global and local metacognitive processing within a specific domain. Similar to anosognosia, functional cognitive disorder, a condition characterized by the experience of persistent and distressing subjective cognitive difficulties in the absence of detectable objective cognitive deficit and underlying neurological disease (98,99), is thought to be explained by changes in metacognitive ability. However, it is unknown which layer(s) of the metacognitive hierarchy, if any, are affected in this condition. Likewise, patients with motor conversion disorder report difficulties in performing certain motor actions without
any apparent neurological disease. Prior work using a visuomotor task revealed that patients are just as aware and confident in trajectory deviations as control participants, but they engaged distinct brain networks when estimating their confidence (100). In this case, distortions in the formation of global self-beliefs may be central in explaining a mismatch between an internal subjective experience of poor self-ability and otherwise intact objective performance and local metacognition (Figure 3).

The various layers in a putative metacognitive hierarchy are likely to be more fine-grained than the local/global dichotomy highlighted here. For instance, we can make a distinction between “how well did I perform this task today at work?” and “how well am I performing at my job in general?”. The levels of metacognition outlined here (Figure 3) partly map onto a previously proposed psychological framework for characterizing global awareness in dementia (101) that distinguishes four levels: sensory pre-registration (basic evaluation), performance monitoring (corresponding to so-called local metacognition here), evaluative judgement and meta-representation. However, in this model, the latter two constructs were defined in relation to how others see us, rather than in relation to objective experimental measurements.

Interim conclusion
Building a complete theoretical framework supported by empirical evidence of how various levels of metacognition relate to each other is important since global self-beliefs are a major determinant of our behavior. Unlike local metacognition, which is often tied to a particular task or cognitive domain, changes in global self-beliefs may generalize to other domains and to a range of daily life functions (89). In turn, global self-beliefs may be more directly relevant for understanding the mechanistic and computational bases of global aspects of subjective experience such as low mood or self-esteem characteristic of negative schemas in depression (80).

Section 3: A transdiagnostic approach for uncovering associations between metacognition and mental health symptoms
If local and global metacognition are to be neurocognitive markers for psychopathology, their robustness and specificity are important. Psychiatric research suggests that the use of the Diagnostic and Statistical Manual of Mental Disorders (DSM) categories pose a concern for these goals (102) due to high comorbidity rates, and, symptom variability
and complexity within each diagnosis (Figure 4A & 4B). For instance, a reduction in memory confidence is often observed in OCD individuals but this has been linked to elevated levels of other mental health symptoms in OCD patient samples (e.g., depression), rather than obsessive-compulsive symptoms per se (22). Hence, accounting for co-morbid symptoms appears crucial for understanding the precise clinical consequences of abnormalities in metacognition, and ultimately allow us to map symptoms more closely to behavior and neural circuits (102–104) (Figure 4C).

Transdiagnostic studies of local metacognition
Recent studies have leveraged transdiagnostic approaches to uncover links between symptom dimensions and metacognition. With self-reported symptoms in nine psychiatric questionnaires (105), we characterized large online general population samples along three symptom dimensions (anxious-depression, compulsive behavior and intrusive thought (henceforth ‘compulsivity’) and social withdrawal; replicated from a prior study (106)). Using a perceptual decision-making task and local confidence ratings, we found that the anxious-depression dimension was associated with lower confidence, whereas the compulsivity dimension was related to higher confidence (Figure 5). These results stand in contrast to classic questionnaire scores showing that OCD symptoms alone were not linked to any alterations in confidence (Figure 5), similar to prior findings (107,108). This is because anxiety and depression, which are both linked to lower local confidence judgments, overlapped with OCD scores (109,110), masking a positive association between confidence and compulsivity. These findings suggest that metacognitive dysfunctions previously observed may be masked by the co-occurrence of other symptoms, particularly if different families of symptoms predict opposing effects on confidence.

A transdiagnostic approach therefore provides context for interpreting prior metacognition findings in case-control studies of OCD. Vaghi and colleagues employed a reinforcement learning task where participants predicted where a particle will land and report their confidence in catching the particle (108). They observed a form of decreased metacognitive sensitivity in OCD as compared to healthy participants (smaller correlation between confidence and behavioral adjustments of their prediction), without a difference in local confidence or in how sensitive participants’ confidence was to task events (e.g., sudden changes in landing location). Conducting the same paradigm in a large online
general population sample, we replicated Vaghi et al.’s finding of an impaired relationship between confidence and behavioral adjustments in OCD (111). However, using a dimensional approach, we found that higher confidence (as in the perceptual task (110)) (Figure 5) and a lower sensitivity of confidence to task events were linked to compulsivity symptoms. These studies demonstrate that transdiagnostic approaches can be crucial in delineating hidden metacognitive relationships and enhancing our understanding of psychopathology.

To our knowledge, the transdiagnostic studies presented above are the only ones applying such approaches to local metacognitive metrics. By using the same three-dimensional structure across multiple studies, we can prevent the overfitting of new psychiatric dimensions to data. Indeed, the same compulsivity dimension linked to metacognitive deficits (105,111), is also associated with goal-directed failures (106), enhanced learning from safety than threat (112), reduced avoidance of cognitive effort (113) and faulty neural representations of task structure knowledge (114). In the case of goal-directed control, deficits are seen in online (106) and in-person samples (114) alike and work in patients has shown these deficits are more strongly linked to variation in a compulsive dimension than a diagnosis of OCD (115). Although these findings are suggestive, it remains to be seen if the metacognitive abnormalities associated with these dimensions are also altered in patient samples. We also note that these dimensions may not necessarily describe cognitive alterations better than DSM-defined psychopathology or other transdiagnostic structures (116–118). Alternative dimensional or hierarchical approaches to phenotyping (119) remain to be tested in the context of metacognition, and may be superior (120–122). As psychiatry continues to improve how we define mental health and illness in the population, we can expect cycles of iterative evolution of dimensional phenotypes (both those of interest and those to be controlled for) (123).

**Intersecting hierarchies of metacognition with transdiagnostic approaches**

Transdiagnostic approaches have revealed that individuals with strong anxious-depression symptoms have lower local confidence, whereas those with compulsivity have higher confidence (105,111). However, the same individual can experience both anxiety and compulsivity symptoms (e.g., OCD). We argue that such opposing effects of confidence between anxious-depression and compulsivity may be unraveled by better
distinguishing between local confidence and global self-beliefs. It is likely that an individual’s local belief about performance is not pure and instead involves numerous, and at least partially dissociable, neural and computational processes. Local confidence ratings in anxious-depression may be ‘contaminated’, i.e., driven by global estimates of self-performance unrelated to the current task, while local confidence ratings in compulsivity could reflect selective abnormalities in local evidence evaluation processes. This explanation is supported by observations that anxiety and depression symptoms are strongly linked to low self-esteem (72,73) while compulsivity is associated with difficulties in developing and using models to solve decision-making tasks (114,124). In sum, a local confidence rating could depend both on a global prior about self-ability and a local evaluation of performance.

Schizophrenic patients have frequently been reported to be overconfident about individual (local) decisions (2,37). However, recent moderation analyses suggest that this metacognitive deficit is based on studies in which other cognitive performance features vary across participants, thereby questioning whether the overconfidence effect is a central deficit (32). This issue is likely exacerbated by the inclusion of variable diagnoses (e.g., bipolar disorder or depression with psychosis) beyond schizophrenia in prior studies (32). Certain forms of schizophrenia also include high levels of apathy which could be partly linked to low global subjective expectations of success (125). As positive and negative symptoms co-exist in schizophrenia, combining a transdiagnostic perspective while considering different levels of metacognition may be fundamental to delineating the underlying psychopathology. For this reason we advocate that future studies use tasks that can distinguish, and simultaneously control for, multiple levels of metacognition (80). Cross-task comparisons might prove useful too, as we hypothesize that reductions in local confidence in depression, if driven by global self-beliefs, should be relatively impervious to task design, and generalize across domains (10). In contrast, if local confidence biases in compulsivity are the result of an issue with ‘model-building’, we expect the finding of over-confidence to be highly sensitive to task demands.

Clinical implications
Metacognitive beliefs have long been a therapeutic target. Metacognitive therapy (MCT) for anxiety, depression (70,126,127), OCD (128,129) and schizophrenia (130,131) focus on modifying intrusive thoughts and cognitive biases to dampen maladaptive rumination,
compulsive rituals or delusional ideation. However, MCT efficacy is not useful for all patients (132–134), and little is known about the underlying neural mechanisms facilitating symptom alleviation (135). Assessing metacognition before and after MCT treatment should help formalize a mechanistic and neural model of how clinical gains occur, and establish if it is through metacognitive processes. Meanwhile, recent studies have shown that training can improve metacognitive ability (136,137) (though with exceptions (138)). A next step is to examine if these metacognitive changes have therapeutic benefit, i.e., transfer beyond a particular training or therapeutic session and generalize to real-world functioning. Gaining an understanding of the factors promoting generalization will be critical for devising tools for improving metacognition (136,137,139) and modifying self-beliefs through psychotherapy (70,140).

The current evidence for a relationship between mental health and metacognition is correlational. Translating these insights to the clinic requires probing these associations causally and in longitudinal designs. A key question is whether abnormalities in metacognitive bias and sensitivity resolve when symptoms improve, or are relatively stable traits that may signal an overall risk for developing a mental health condition. Drawing on adjacent literature, there is some evidence to suggest that negative biases in face perception improve following antidepressant drug administration in depressed patients and predict subsequent clinical response (141). If metacognitive bias follows a similar pattern as negative biases, it may constitute a similar predictor of treatment outcome. Quantifying metacognition could therefore have clinical value if changes in metacognitive parameters help to identify individuals at risk, facilitate early intervention, guide us as to who might respond best to a given treatment, or assist in developing transdiagnostic treatment protocols that target metacognition (142–144).

CONCLUSIONS
Theories about the role of metacognition in mental health may be enriched by adopting quantitative task-based methods for measuring metacognition across different hierarchical levels (Figure 3) together with robust transdiagnostic approaches (Figure 4). Many other aspects of metacognition have yet to be looked at in relation to mental health, and the paradigms and models described here represent a starting point. The current review serves as a framework for thinking about how different levels of metacognition (from local to global) are interrelated, possibly by generalization.
mechanisms, and outlines hypotheses for how these map onto transdiagnostic dimensions of mental health.

ACKNOWLEDGEMENTS
TXFS is a post-doctoral fellow at the Max Planck UCL Centre for Computational Psychiatry and Ageing Research. The Max Planck UCL Centre for Computational Psychiatry and Ageing Research is a joint initiative supported by UCL and the Max Planck Society. The Wellcome Centre for Human Neuroimaging is supported by core funding from the Wellcome Trust (203147/Z/16/Z).

MR is the beneficiary of a post-doctoral fellowship from the AXA Research Fund. MR work was also supported by a department-wide grant from the Agence Nationale de la Recherche (ANR-17-EURE-0017, EUR FrontCog). This work has received support under the program «Investissements d’Avenir» launched by the French Government and implemented by ANR (ANR-10-IDEX-0001-02 PSL).

CMG is supported by a fellowship from MQ: transforming mental health (MQ16IP13) and holds grant funding from Science Foundation Ireland’s Frontiers for the Future Award (19/FFP/6418).

SMF is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and Royal Society (206648/Z/17/Z).

For the purpose of Open Access, the authors have applied a CC-BY public copyright licence to any author accepted manuscript version arising from this submission.

CONFLICTS OF INTEREST
The authors report no biomedical financial interests or potential conflicts of interest.
FIGURE LEGENDS

Figure 1. Experimental task-based measures of metacognition.
A) Relationship between first-order performance and second-order confidence (“local” trial-by-trial confidence judgements). On each trial, participants provide a report of their level of confidence in a decision/choice they have made, which can either be objectively correct or incorrect.
B) Two independent metrics of metacognition - bias and sensitivity. Each schematic graph shows a probability distribution of confidence ratings for correct and incorrect trials separately. The x-axis represents confidence reports increasing from left to right. Metacognitive sensitivity is the extent to which confidence discriminates between correct and incorrect trials, corresponding to the separation between the distributions. Metacognitive bias is the overall confidence level across both correct and incorrect trials. Confidence distributions are Gaussian for illustration purposes but are likely to take other forms depending on the generative model.

Figure 2. Neural correlates of metacognitive evaluation
Schematic sagittal slice and lateral view of the human brain highlighting the role of prefrontal cortex (PFC) in metacognition. Studies of local metacognition have highlighted the ventro-medial prefrontal cortices (vmPFC) and posterior medial frontal cortex (pMFC) as central hubs reflecting confidence estimates [a: (9,145)] and error detection [b: (146–148)], while the frontopolar cortex (FPC), together with the lateral prefrontal cortex (lPFC) are involved in mediating explicit metacognitive judgements, (meta)cognitive control and subsequent behavioral regulation [c: (8,47); d: (40,43)]. Some of the neural substrates linked to local metacognition exhibit cognitive domain-specificity e.g., the precuneus (PRECU) has mostly been implicated in metamemory [e: (9,44,46,149)], whilst lateral-parietal areas (lPAR) are mostly implicated in metaperception [f: (47,150)]. Recent studies have begun to reveal that neural substrates of global metacognition only partly overlap with those of local metacognition. In particular, in vmPFC and precuneus, local confidence signals were found to be further modulated by the level of global self-belief on a perceptual task (81).

Figure 3. Multiple hierarchical levels of metacognitive evaluation
Reciprocal interactions between “local” confidence judgements in isolated decisions and more global self-beliefs. Previous work has revealed that local confidence contributes to the formation of global self-beliefs, but global self-beliefs are also likely to in turn influence local confidence. Under this framework, local confidence may reflect a combination of a local component related to decision performance evaluation and a global component formed over the aggregation of multiple experiences across various tasks and domains formed through learning. On the right, examples are given to illustrate each hierarchical level in the domain of memory, though the true distinction between levels is likely to be more fine-grained. Each of these metacognitive levels is associated with dynamics unfolding across different timescales, with higher levels of the
hierarchy having slower dynamics than lower levels. Global self-beliefs may shape and be shaped by even more global constructs such as an individual’s level of self-esteem.

**Figure 4.** Dimensional approaches to psychiatry addressing within- and across-diagnosis homogeneities and heterogeneities. 

A & B) Case-control studies comparing diagnosed patient and healthy control groups have often failed to recognize that patients have varying levels of other psychopathologies (e.g., compulsivity, anxiety, etc.) beyond the one under study. Comparing such groups (typically, ranging between 15 and 50 participants per group) have often revealed ambiguous or non-specific effects in relation to metacognition.

C) Mathematical methods of dimensionality reduction allow identification of latent factors underlying various mental health conditions. These dimensions may better reflect the psychopathological complexity underlying traditional psychiatric categories, and uncover more consistent relationships with metacognition. OCD (obsessive-compulsive disorder) and GAD (generalized anxiety disorder) reflect traditional diagnostic categories. In contrast, Anxious and Compulsive dimensions reflect transdiagnostic symptom dimensions. Typically, transdiagnostic dimensions are estimated using groups of hundreds or thousands of participants.

**Figure 5.** Relationships of confidence and psychiatric symptoms (standard approach), or with psychiatric dimensions (transdiagnostic approach), across two different paradigms. AD: anxious-depression dimension, CIT: compulsive behavior and intrusive thought (“compulsivity”) dimension, SW: social withdrawal dimension. Confidence abnormalities linked to psychiatric symptoms using the standard approach are inconsistent across studies. However, with a transdiagnostic approach, the finding of lowered confidence with anxious-depression and higher confidence with compulsivity replicates across tasks. The y axis indicates the change in z-scored confidence for each change of 1 standard deviation of symptom/dimension scores. Error bars denote standard error. *p < .05, **p < .01, ***p < .001 corrected for multiple comparisons, °p < .05, uncorrected. Figures are reproduced from their original studies (105,111). Note that performance was controlled for using a staircase procedure in the perceptual discrimination task and was not related to symptom dimensions (105). Task performance also showed no relationship with symptom dimensions in the reinforcement learning task (111).
REFERENCES

1. Flavell JH. Metacognition and cognitive monitoring: A new area of cognitive–


37. Moritz S, Woodward TS, Rodriguez-Raecke R. Patients with schizophrenia do not produce more false memories than controls but are more confident in them. Psychol Med. 2006;36(5):659.
41. Hilgenstock R, Weiss T, Witte OW. You’d better think twice: post-decision


62. Haddara N, Rahnev D. The impact of feedback on perceptual decision making and metacognition: Reduction in bias but no change in sensitivity. 2020;


84. Lee ALF, de Gardelle V, Mamassian P. Global visual confidence. 2020;
96. Mazancieux A, Souchay C, Casez O, Moulin CJA. Metacognition and self-


114. Seow TXF, O’Connell R, Gillan CM. Model-based learning deficits in compulsivity are linked to faulty representations of task structure. bioRxiv [Internet]. 2020; Available from: https://doi.org/10.1101/2020.06.11.147447


119. Dalgleish T, Black M, Johnston D, Bevan A. Transdiagnostic approaches to


124. Sharp PB, Dolan RJ, Eldar E. Disrupted state transition learning as a computational marker of compulsivity and anxious arousal. 2020;


2019;101751.


Figure A: 
- First-order Performance
- Second-order Confidence
- Correct, Incorrect, Correct
- Very confident, Unconfident, Unsure
- Trials

Figure B: 
- Metacognitive Sensitivity
  - Low: incorrect, correct
  - High: incorrect, correct
- Metacognitive Bias
  - Low: low confidence, high confidence
  - High: low confidence, high confidence
Explicit metacognitive reports [c]

Use of metacognitive representations for implementing cognitive control [d]

Domain-general hub for error detection [b]

Domain-general hub for confidence formation [a]

Tracking of mnemonic evidence [e]

Tracking of sensory evidence [f]

Areas involved in:
- local (decision) confidence
- local (decision) and global (task) confidence
A hierarchy of metacognitive evaluation

- Cognitive domains
- Tasks
- Decisions (trials)

Global
- «I generally have good memory»

Local
- «I did pretty well on this exam»
- «I got this question correct»
**Perceptual discrimination task**

**Reinforcement learning task**

---

- Standard
- Transdiagnostic

**Contribution to Confidence**

- Alc. Addiction
- Apathy
- Depression Disorder
- Impulsivity
- OCD
- Schizotypy
- Social Anxiety
- Trait Anxiety

Legend:
- *******: p < 0.001
- ****: p < 0.01
- *: p < 0.05