ROUTLEDGE ENCYCLOPEDIA OF PHILOSOPHY

Metacognition

Joëlle Proust, Institut Jean-Nicod (Paris)

Keywords

Uncertainty, confidence, reliability, mental action, calibration, noetic feelings, appraisal.

Summary

It has often been claimed that metacognition should be defined as 'cognition about one's own cognition', 'knowledge about one's own knowledge', or 'thinking about one's own thinking' (Carruthers, 2011, Nelson & Narens, 1992, Perner, 2012). These formulations, however, are now often seen as unduly restricting the scope of metacognition to a form of reflective judgment. There is evidence that agents with no concept of perception or knowledge, such as monkeys and young children, are nevertheless able to assess when they can confidently engage in a task (such as finding an object, discriminating visual patterns or recognizing whether an item was already presented) (Hampton, 2009). On an alternative definition, then, metacognition is the ability to evaluate whether one is likely to achieve a specific cognitive goal or to have successfully achieved it. These evaluations jointly contribute to the control of one's actions. They allow agents to select contextually efficient cognitive actions, such as trying or not trying to remember the location of an object , and to decide whether or not to rely on a specific cognitive output to act on the world. In both cases, on this broader definition, agents may rely either on the degree of confidence that they experience (a noetic feeling), or on what they judge to be the case, given their background beliefs and concept-based predictions (Arango-Muñoz 2011, Koriat & Levy-Sadot 1999, Proust 2007, 2013, 2015). Metacognition so conceived is distributed in domain-specific abilities, involved in the regulation of mental actions, the distribution of resources between rival processes, and the regulation of one's own emotions. Inter-agent metacognitive control is also exemplified in the pragmatic principles of conversation first explored by Paul Grice, and in collective forms of epistemic investigation and decision-making (such as in scientific research).

Defining Metacognition

Metacognition is best defined through its functional role within cognition (Shea et al. 2015). 'Cognition' extends to all the informational processes involved in representing the world and acting on it, with functions such as perceiving, desiring, learning, remembering, reasoning, intending and acting. 'Metacognition' refers to the specific processes whose function is a) to control the informational reliability of the cognitive activity in these various dimensions and b) to monitor the informational reliability of the processes engaged when performing specific cognitive tasks. What is meant by 'informational reliability' is the probability of information to be adequately channelled, retrieved and combined in generating decisions. Such decisions are involved in selecting or monitoring cognitive actions, i.e. actions performed in order to acquire or extend one's own knowledge, such as trying to discriminate, remember, or solve a problem (Arango-Muñoz, S. 2014, Proust, 2013). Control is related to the rational *selection of a given cognitive action* as a function of the agents' present goals and available resources (in particular, the amount of cognitive effort to be allocated for the action to succeed). Monitoring consists in *assessing the informational adequacy* of the current performance, by tracking deviations from the epistemic goal (perceptual or reasoning error, memory failure, inconsistency in beliefs or in preferences, improper concept use, etc.).

The defining property of metacognition does not consist in the specific control mechanisms that are being used: control and monitoring mechanisms are ubiquitous at all levels of behaviour (Carruthers, 2015, Shea et al. 2015). Furthermore, the mechanisms relevant to metacognition, while important for theorizing about metacognition, are only mentioned in its definition for their function: selecting appropriate informational goals and assessing the validity of particular decisions in the domain of perception, memory, or reasoning. What does it mean, however, to claim that a metacognitive sybsystem "tracks the informational reliability of first-order cognitive actions"?, To clarify this claim, a primary distinction between subjective and objective uncertainty needs to be made.

Two types of uncertainty

As observed by David Hume (*Treatise*, I, 4, 1), our ability to form true beliefs about the world is threatened by two forms of uncertainty. Objective uncertainty originates in the evidence on which our judgments are built – i.e. on the way the world appears to us. The more stable the world, the more likely it is that a prediction about it will turn out to be correct. A perpetually changing world would not be knowable, because information crucially depends on the existence of causal regularities (Dretske, 1988). Subjective uncertainty originates in the agent's own variable cognitive dispositions. Subjective probability of error can be assessed, however, through "a reflex act of the mind, wherein the nature of our understanding and our reasoning from the first probability become our objects" (*Treatise*, p. 122). A 'reflex act of the mind' is one through which the mind assesses how likely it is that it correctly represents a given state of affairs. For example, recognizing an apparently familiar face in an unexpected location leads to doubting one's own perception.

A sceptical argument about metacognition

David Hume considers the sceptical move through which our primary *subjective* uncertainty (did I actually recognize N?) is seen as applicable to our critical estimation itself (am I right to think that I did ?). This recursive process, iterated in infinity, should progressively diminish the probability for a belief to be correct toward a "total extinction of belief and evidence" (Hume, 2007, p. 122). Hume recognizes the cogency of the argument, but describes our beliefs as being de facto insensitive to this recursive threat. Such insensitivity indicates, on his view, that reasoning and belief are not governed by ideas and reflections, but rather by "some sensation or peculiar manner of conception". An argument that requires an effort of thought, even when perfectly comprehended, may look far less persuasive than a "lively conception", i.e. a consideration that is easy to represent (Hume, 2007, p. 123). In agreement with Hume, psychological studies indeed suggest that laypersons take 'ease of reasoning' – i.e. 'fluency' – as an indicator of truth (Reber & Unkelbach, 2010). This finding, however, should not suggest that metacognition is merely based on a least effort principle, nor that it tends to motivate irrational decisions. To summarize: Hume's sceptical argument about metacognition was that, when assessments of subjective uncertainty are recursively applied to one's own prior assessments, the subjective likelihood for being correct should tend to zero. As Hume observes, the fact that this argument does not describe humans' way of assessing their uncertainty suggests that metacognition is based on feelings rather than reasons.

Objective and subjective informational reliability

Just as there are two forms of uncertainty, subjective and objective, there are two forms of informational reliability. Objective informational reliability refers to the causal properties of events that makes them objectively predictable (for example the time of the day objectively predicts light intensity). Subjective informational reliability refers to the causal properties of cognitive and metacognitive processing that make errors predictable or detectable (for example slow cognitive processing tends to predict error). In both cases, reliability is measured by the frequency of predictive success. Metacognition assesses subjective informational reliability, that is: the degree of subjective certainty attached to a given task or outcome.

For an agent to generate correct results in a given cognitive task, informational reliability in the first order task (such as remembering) should be consistent with the subjective probabilities that are collected. Calibration mechanisms are important informational channels allowing felt confidence to track over time the objective probability of subjective predictions being correct in a type of cognitive performance. For example, having repeatedly failed to retrieve a proper name from memory leads one to experience a lower confidence in being able to retrieve another proper name. Calibration of confidence is optimal when confidence in a set of performances coincides with objective accuracy in that set. For example, someone with a moderately high confidence in her memory for names is well calibrated if her retrieval rate is about 75% of the names searched, not if her retrieval rate is about 30%. Successful calibration of subjective assessments of confidence is subjected to two conditions. First, cognitive actions of the same type have to be repeatedly performed for collecting sufficient calibrating feedback. Second, the feedback so collected must be informative about action success. In other words, it must not be positively or negatively biased (Fleming et al., 2012). In a positive bias, such as the equality bias, the subjects' decisions are made to appear systematically better than they actually are; in a negative bias, such as the gender bias, they are systematically presented as worse. Overconfidence or underconfidence in performing this type of task results from such biasing effects, often motivated by social prejudices.

A computational approach to metacognition can be represented as addressing the sceptical argument in a more principled way than Hume's proposal. Hume's sceptical argument was that, when assessments of subjective uncertainty are recursively applied to one's own prior assessments, the subjective likelihood for being correct should tend to zero. If subjective uncertainty, i.e., confidence in one's own perceptual discriminations, is modelled as the ratio of the mean of the sensory estimate to its standard deviation, there is no additional deviation to be expected at a higher level. Further calibration allows convergence to occur toward an internal reliability standard. The reason for this convergence is that calibration mechanisms permanently adjust the threshold of confidence to the updated standard deviation (Bahrami et al., 2012).

A sceptic could argue, however, that a calibration mechanism may occasionally fail to be reliable. Social biases, such as an equality bias, a gender bias, or a misguided deference to pseudo-experts, have been demonstrated to promote biased metacognitive feedback (i.e. inadequate feelings of confidence), which jeopardizes calibration (Mahmoodi et al., 2015).

Theories of metacognition: attributivism vs evaluativism

What are the abilities that allow metacognition to develop? A major controversy concerns the scope and informational basis of metacognitive assessments. A number of the early theorists of metacognition, including psychologists John Flavell (Flavell, 1979), Thomas Nelson and Louis Narens, (Nelson & Narens 1992) have claimed that metacognition requires modelling first-order cognitive states, and that, consequently, a form of mindreading is involved in assessing the informational reliability of one's own mental states. Because metacognition is seen as engaging mindreading, i.e., mental state attribution, this view has been referred to as attributivism (Proust, 2013). More recently, attributivism has been further defended by Peter Carruthers (2009, 2011). On this view, non-mindreaders, such as rodents and monkeys, are unable to perform metacognitive control and monitoring of their own mental states.

In contrast, evaluativism is the view that two informational sources are available to human agents in order to predict cognitive accuracy, consisting respectively in emotional experience and in concept possession. The first, which is shared with some non-human species, relies on dedicated emotions, called 'noetic feelings'. For example, one can feel that one knows or does not know a given proper name, while failing to retrieve it; feeling uncertain does not depend on a conceptual competence or a judgment that one can remember a specific name any more than feeling angry depends on having a concept of anger. The other relies on innate or acquired knowledge about mental functions. For example, one may judge that one can respond to a question because it belongs to one's own area of expertise. In such cases, one uses concepts and theories (about learning and remembering) to assess one's own competence.

Empirical arguments have been offered by evaluativists, documenting the contrast between two metacognitive systems. Nonhumans, such as monkeys and rodents, in particular, seem to be unable to judge that they can perceive or remember an item because they do not have the associated mental concepts available. However, they are able to control and monitor their own cognitive activity just as humans do, in tasks involving memory of former presentations, or perceptual discriminations of displays (Call 2010, Couchman et al. 2012, Kornell et al. (2007). Three year-old children, similarly, are notoriously unable to reliably attribute knowledge to themselves and to others, when tested verbally (Gopnik & Astington, 1988). They are able, however, to reliably decide when to inform, or not to inform, another person about objects and properties as a function of what they know to be the case, well before they are able to reliably apply the concept of knowledge to the informational states they have (Kim et al., 2016). Twenty-month old infants have even been demonstrated to ask for help – and hence, communicate their uncertainty to others – only when objectively needed (Goupil et al., 2016): hence, monitoring their own uncertainty does not seem to require from the children that they *judge that they know,* (when 'know' is taken to refer to a mental concept possessed by the child). Finally, this dissociation is also present in human adults; subjects tend to make different predictions about future recall when having engaged themselves in a learning task— a cognitive action— and when having merely observed others perform the task (Koriat & Ackerman, 2010).

Because they define metacognition as requiring a metarepresentation of one's own mental states as mental, the mindreading theorists of metacognition have been unconvinced by these arguments. For example, developmental psychologist Josef Perner (2012, p.113-115), thinks that "apes' desire for getting good looks" (to locate where the food is hidden) does not qualify as metacognitive, "because the animal does not need to know that it lacks sufficient knowledge". This objection amounts to claiming that a conceptual understanding of one's own ignorance is a presupposition of metacognition. Theory-theorists have also emphasized that metacognition in children flourishes well after they are able to pass false-belief tasks. This argument however, has been contradicted by evidence for non-verbal metacognitive decisions in infants and young children (Goupil et al., 2016, Balcomb and Gerken, 2008).

Others, such as Peter Carruthers and J. Brendan Ritchie (2012), recognize that feelings can guide uncertainty-based decision-making, and that evidence for this ability in animals is impressive. They deny, however, that this function is metacognitive (for reasons that, again, are terminological). Animals' decisions to opt out from a cognitive task are seen to belong to the regulation of action, rather than to the regulation of informational reliability. An additional objection to the infant studies stems from modularism (nativism about mindreading) (Carruthers & Ritchie, 2012): children's failure in verbal tests might reflect executive rather than conceptual difficulties. Seen from this viewpoint, dissociation between practical decisions to inform and self-knowledge report does not speak in favor of a two-system view of metacognition.

Noetic feelings

Granting that emotional mental events called 'noetic feelings' play a central role in metacognition, how do we characterize them? Feelings of confidence, of knowing, of being right, and the tip of the tongue phenomenon (experienced when failing to recall a word while sensing that retrieval is imminent) are seen as appraisals of likely success or failure of one's own current cognitive actions (De Sousa, 2009). Variations in valence (pleasurable or aversive) and in intensity guide epistemic decision in an immediate way (Hookway, 2003). These variations have been shown to be based on the discrepancies between observed and expected feedback predicting likely success in a current cognitive task (Proust, 2015a). A noetic feeling of uncertainty of a given degree, for example, motivates agents to try harder to perceive, remember or solve a problem, or to give up their current task. Attributing a crucial evaluative role to feelings in epistemic decision substantially modifies the traditional view that emotions are generally obstacles rather than instruments of rationality (de Sousa, 2009).

There are also controversies about whether noetic feelings have intrinsic or derived intentional content. While some theorists, following Tye (2009), take them to be introspective experiences of first-order states, others hold that they are rather bodily experiences with a merely derived intentional content. On this view, noetic feelings carry information about what might easily happen. Hence, they provide agents with modal knowledge about their own competence in a current task (Dokic, 2012). In contrast, some philosophers take noetic feelings to be representations of intrinsic intentional states. What they transparently indicate, however, is neither a fact about the environment, nor a fact about mental properties, but rather a subjective relation between agents and environments. The intentionality of emotional states, then, is seen as 'Janus-faced' (de Sousa, 2009), or claimed to exemplify 'Pushmi-Pullyu' representations of epistemic affordances (Millikan, 1995) endowed with a non-propositional, associative format (Proust, 2015b).

Bibliography

Arango-Muñoz, S. (2011). Two levels of metacognition. *Philosophia*, *39*(1), 71-82. DOI: 10.1007/s11406-010-9279-0. (presents the standardly accepted view that metacognition includes both experience-based and concept-based evaluations.)

Arango-Muñoz, S. (2014). The nature of epistemic feelings. *Philosophical Psychology*, *27*(2), 193-211. DOI:10.1080/09515089.2012.732002.(studies the role of epistemic feelings in metacognition)

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., & Frith, C. (2012). What failure in collective decision-making tells us about metacognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1350-1365. DOI: 10.1098/rstb.2011.0420. (shows in which conditions exchanging opinions about cognitive matters enhances the validity of the resulting collective judgment)

Balcomb, F. K. and Gerken, L. (2008). Three-year old children can access their own memory to guide responses on a visual matching task. *Developmental Science*, 11(5), 750–60. (the first study showing that children can evaluate what they remember before being able to solve a false belief task).

Call, J. (2010). Do apes know that they could be wrong?. *Animal cognition*, *13*(5), 689-700. DOI:10.1007/s10071-010-0317-x. An experimental study of the search behavior in chimpanzees shows that the animals are sensitive to the amount of information that they have.

Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and brain sciences*, *32*(02), 121-138. DOI: 10.1017/S0140525X09000545. This article discusses four different accounts of the relations between mindreading and metacognition.

Carruthers, P. (2011). *The Opacity of Mind: an integrative theory of self-knowledge*. Oxford: Oxford University Press. This book is about the nature and sources of self-knowledge, particularly on our knowledge of our current thought processes.

Carruthers, P., & Ritchie, J. B. (2012). The emergence of metacognition: affect and uncertainty in animals. In Beran, M.J., Brandl, J., Perner, J. & Proust, J. (Eds.), *Foundations of Metacognition*, Oxford: Oxford University Press, 76-93.
This article defends that although animals guide their decisions to act through feelings, these feelings are not specifically metacognitive.

Couchman, J.J., Beran, M.J., Coutinho, M.V.C., Boomer, J. & Smith, J.D. (2012). Evidence for animal metaminds. In Beran, M.J., Brandl, J., Perner, J. & Proust, J. (eds.), *Foundations Metacognition*, Oxford: Oxford University Press, 21-35.
This chapter reviews comparative evidence demonstrating metacognition in nonhumans.

De Sousa, R. (2009). Epistemic feelings. *Mind and Matter*, *7*(2), 139-161.
This article defends the view that epistemic feelings include specialized variants of fear and greed.

Dokic, J. (2012). Seeds of self-knowledge: noetic feelings and metacognition. In Beran et al. (eds.), *Fondations of Metacognition,* 302-321. Oxford: Oxford University Press.
This chapter proposes that noetic feelings are about one's own cognitive competence, but do not need to involve metarepresentational abilities.

Dretske, F. I. (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge, MA: MIT press. This book proposes a compelling account of the nature and causal role in action guidance of intentional content.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive developmental inquiry. American Psychologist 34 (1979): 906-911.
This article claims that metacognition includes four types of phenomena: knowledge, experience, goals and strategies. It is incompatible with the existence of non-human or human concept-free metacognition.

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: computation, biology and function. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *367*(1594), 1280-1286. DOI: 10.1098/rstb.2012.0021.

This article presents experimental paradigms for analysing metacognition with neuroscientific and computational methods and speculates about its function.

Goupil, L., Romand-Monnier, M., & Kouider, S. (2016). Infants ask for help when they know they don't know. *Proceedings of the National Academy of Sciences*, *113*(13), 3492-3496.
This study shows that 12- and 18-month-old preverbal infants are able to internally monitor the accuracy of their own decisions.

Hampton, R. R. (2009). Multiple demonstrations of metacognition in nonhumans: Converging evidence or multiple mechanisms?. *Comparative cognition & behavior reviews*, *4*, 17-28.
A review of the comparative literature suggests that the demonstrations of metacognition in nonhumans can be explained in terms of associative learning or other mechanisms that do not require invoking introspection or access to private mental states.

Hookway, C. (2003). Affective states and epistemic immediacy. *Metaphilosophy*, vol.34: 78-96. Reprinted in (2003). Michael Brady and Duncan Pritchard (eds), *Moral and Epistemic Virtues*, Oxford, Blackwell: 75-92.
The author demonstrates the epistemological role of emotions in epistemic evaluation and invokes traits of character in order to explain how these affective evaluations are regulated.

Hume, D. (1888, 2007). *Treatise on Human Nature,* D. Morton & M. Norton (eds.), Oxford, Clarendon Press.
This book attempts to explain the principles of human nature on the basis of an analysis of understanding, of the nature of our ideas, and of our reasoning processes

Kim, S., Paulus, M., Sodian, B., & Proust, J. (2016). Young Children's Sensitivity to Their Own Ignorance in Informing Others. *PloS one*, *11*(3), e0152595.
This article demonstrates that children' evaluations of what they know differ when their goal is to tell whether or not they know that *P* or to inform another person about *P*.

Koriat, A., & Ackerman, R. (2010). Metacognition and mindreading: Judgments of learning for self and other during self-paced study. *Consciousness and cognition*, *19*(1), 251-264. doi:10.1016/j.concog.2009.12.010.
This study shows that people do not use cognitive effort to predict success in learning in the same way when they have performed a task or merely observed some one else perform it.

Koriat, A., & Levy-Sadot, R. (1999). Processes underlying metacognitive judgments: Information-based and experience-based monitoring of one's own knowledge. In: S. Chaïken & Y. Trope, (Eds.), *Dual-Process Theories in Social Psychology*. London: The Guilford Press, 483-502.
This chapter offers a systematic defence of the specific contribution of experience-based and concept-based metacognition.
Kornell, N., Son, L. K., & Terrace, H. S. (2007). Transfer of metacognitive skills and hint seeking in monkeys. *Psychological Science*, *18*(1), 64-71. doi: 10.1111/j.1467-9280.2007.01850.
Rhesus macaques, trained previously to make *retrospective* confidence judgments about their performance on perceptual tasks, transfer that ability immediately to a new task, and can also learn to request "hints" helping them to solve given problems.

Mahmoodi, A., Bang, D., Olsen, K., Zhao, Y. A., Shi, Z., Broberg, K., ... & Roepstorff, A.

(2015). Equality bias impairs collective decision-making across cultures. *Proceedings of the National Academy of Sciences*, *112*(12), 3835-3840. doi: 10.1073/pnas.1421692112. This study has shown that in several cultures, people try to reach a common decision even when they have had unequal access to the relevant information.

Millikan, R. (1995) Pushmi-Pullyu Representations, *Philosophical Perspectives*, 9:185-200.
A classical article in which Millikan observes that some representations seem to have both a belief and a desire relation to the world.

Nelson, T. O. and Narens, L. (1992). Metamemory: a theoretical framework and new findings, in T. O. Nelson (Ed.). *Metacognition, Core Readings.* Boston: Allyn & Bacon, 117-130.
This article proposes a theoretical framework for metamemory research that has exerted a deep influence on the whole field of metacognition.

Perner, J. (2012). MiniMeta: in search of minimal criteria for metacognition. In M. Beran, J. Brandl, J. Perner & J. Proust (eds.), The *Foundations of Metacognition*, Oxford: Oxford University Press: 94-116.
This article raises problems associated with behavioral studies of animal metacognition, and presents criteria that should guide recognizing a decision as involving metacognition

Proust, J. (2007). Metacognition and metarepresentation: is a self-directed theory of mind a precondition for metacognition? *Synthese*, *159*(2), 271-295. DOI: 10.1007/s11229-007-9208-3.
This article proposes a view of metacognition able to account for nonhumans' ability to opt out from difficult tasks, search information, or determine when their response is correct or not.

Proust, J. (2013). *The Philosophy of Metacognition. Mental agency and self-awareness.* Oxford: Oxford University Press.
This book confronts the attributive and the evaluative view of metacognition, and proposes an analysis of the normative structure of cognitive actions.

Proust, J. (2015a). Time and action: Impulsivity, habit, strategy? In *The Review of Philosophy and Psychology 6(4), 717-743.* doi10.1007/s13164-014-0224-1.
This article proposes that there are three types of physical and cognitive action depending on a rational trade-off between time and cognitive resources, on the one hand, and benefits and risks, on the other.

Proust, J. (2015b). The Representational Structure of Feelings, in T. Metzinger & & J.M. Windt, (eds.) *Open Mind.* www.open-mind.net. doi: 10.15502/9783958571044.
This article offers an analysis of the semantic structure that belongs to all kinds of feelings, whether bodily, socio-affective or metacognitive.

Reber, R., & Unkelbach, C. (2010). The epistemic status of processing fluency as source for judgments of truth. *Review of philosophy and psychology*, *1*(4), 563-581. DOI: 10.1007/s13164-010-0039-7.
This article discusses the implications of the view according to which processing fluency is a potentially universal cue to holding true a proposition.

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in cognitive sciences*, *18*(4), 186-193. DOI: http://dx.doi.org/10.1016/j.tics.2014.01.006.
This article proposes that the System-2 type of Metacognition has evolved in order to broadast metacognitive information to others.

Tye, M. (2009). *Consciousness revisited: Materialism without phenomenal concepts*. Cambridge, MA: MIT Press.
This book criticizes the phenomenal-concept strategy for defending materialism about consciousness and presents an alternative view that addresses the traditional puzzles through Russell's distinction between knowledge by acquaintance and knowledge by description.

Further Readings

Carruthers, P. (2008). Meta-cognition in animals: A skeptical look. *Mind and Language, 23,* 58–89.
This article defends that animals are unable to control their cognitive states as humans do.

Recanati, F. (2000). *The Iconicity of Metarepresentations*. In D. Sperber (ed.) *Metarepresentations*. *A Pluridisciplinary perspective*. Oxford: Oxford University Press. pp. 311-360.
This article offers a renewed semantic analysis of metarepresentations emphasizing their iconic and simulative character.