Lecture 04 : Metacognition : Two Systems

Corrado Sinigaglia & Stephen A. Butterfill
< >

Tuesday, 29th March 2022

Contents

1. Metacognitive Feelings: How Do Fast and Slow Processes Interact?

How, if at all, do fast and slow processes influence each other?

We have seen that fast and slow processes can yield incompatible responses to a single scenario (in both mindreading and physical cognition; see *Mindreading: Signature Limits, and Development* in Lecture 02 and *Speed-Accuracy Trade-Offs (in Physical Cognition)* in Lecture 01). This suggests that the representations fast and slow processes operate over are not inferentially integrated.

Because of how we characterised what it is for systems to be distinct, there is a tension between postulating two (or more) systems and postulating interactions between them. We suggested that the distinctness of systems consists in there being processes which differ in conditions which influence whether they occur, and which outputs they generate (in *The Core Idea* in Lecture 01). As the scope for interaction increases, the grounds for distinguishing systems weaken.

Earlier, in *Speed-Accuracy Trade-Offs (in Physical Cognition)* in Lecture 01, we saw that it is possible for a fast process to influence a slow one indirectly and asynchronously if the fast system can modify the overall phenomenal character of experiences. This provides one model for understanding interactions between fast and slow systems.

But is it also possible for a fast process to influence a slow one synchronously?

1.1. Metacognitive Feelings as a Bridge

According to Koriat,

> 'metacognitive feelings … allow a transition from the implicit-automatic mode to the explicit-controlled mode of operation' (Koriat 2000, p. 150).

Koriat's focus is not two-systems theories, but his claim hints that metacognitive feelings might be relevant to understanding how fast processes could influence slow processes.

1.2. What Are Metacognitive Feelings?

Metacognitive feelings include:

- familiarity (Whittlesea & Williams 1998; Scott & Dienes 2008)
- the feeling of knowing (Koriat 2000)
- feeling that a name is on the tip of your tongue (Brown 1991)[1]
- the feeling you have when someone's eyes are boring into your back
- Déjà vu (Brown & Marsh 2010)
- ? surprise (Reisenzein 2000)
- the feeling of being the agent of an event ('sense of agency') (Haggard & Chambon 2012)

This is not supposed to be an exhaustive list. Dokic (2012) lists several more, and others have postulated novel metacognitive feelings (for example, Velasco & Casati (2020) argue that there is a metacognitive feeling of disorientation). It is also possible that some items on the list do not qualify as metacognitive feelings.

What makes something a metacognitive feeling? We adapt an idea from Dokic:

> 'the causal antecedents of [certain] feelings can be said to be metacognitive insofar as they involve implicit monitoring mechanisms that are sensitive to non-intentional properties of first-order cognitive processes' (Dokic 2012, p. 310).

We propose that a metacognitive feeling is a feeling which is caused by a metacognitive process, that is, a process which monitors another cognitive process. For example, a process which monitors the fluency of recall, or of action selection, is a metacognitive process.

1.3. The Feeling of Familiarity

What causes feelings of familiarity? Not familiarity as such, it turns out. Instead they are caused by the ease with which you can process the features of a face relative to difficulty of identifying the person. Roughly, the greater the discrepancy between fluency of processing and difficulty of identification, the stronger the feeling of familiarity (Whittlesea & Williams 1998).

So what is this feeling of familiarity?

First, it is phenomenal. It is an aspect of the phenomenal character of some experience associated with acting. So we can call it a feeling.

[1] Widner et al. (2005) provides evidence that the feeling of knowing is distinct from the feeling that something is on the tip of your tongue.

Second, it is metacognitive in the sense that it's normal causes include processes which monitor fluency of processing. This is why the feeling of familiarity counts as a metacognitive feeling.

Third, it does not necessarily give rise to beliefs. The feeling does not lessen even if you believe (or know) that the thing which causes your feeling of familiarity is not one you have ever encountered before.

Fourth, you are not forced to treat feelings of familiarity as being about actual familiarity: instead you can use feeling of familiarity in deciding whether a stimulus is from that grammar (Wan et al. 2008). In this respect, metacognitive feelings are unlike perceptual experiences and unlike emotions. As Dokic observes:

> 'It is difficult to imagine fear that does not have the function of detecting danger. In contrast, many [metacognitive] feelings seem to be recruited by the organism through some form of learning' (Dokic 2012, p. 308).

1.4. The Sense of Agency

Feelings of agency, seem to arise from a number of cues including comparison between outcomes represented motorically and outcomes detected sensorily and the fluency of an action selection process (that is, the ease or difficulty involved in selecting one among several possible actions to perform motorically; this can be manipulated by, for example, providing helpful or misleading cues to action (Wenke et al. 2010; Sidarus et al. 2013, 2017)).

The sense of agency is relevant to us because it serves to link two largely independent processes concerned with evaluating whether you are the agent of an event. One involves detecting the cues just mentioned; the other involves thinking about how likely it is that you are the agent of an event, perhaps in the light of your background knowledge.

1.5. Surprise

Are there metacognitive feelings of surprise?

> 'the intensity of felt surprise is not only influenced by the unexpectedness of the surprising event, but also by the degree of the event's interference with ongoing mental activity, [...] the effect of unexpectedness on surprise is [...] partly mediated by mental interference' (Reisenzein 2000, p. 271).

That is, there is a feeling of surprise which is a sensational consequence of mental interference. (This can be tested by increasing cognitive load: this

intensifies feelings of surprise without, of course, making the events themselves more surprising. But see Reisenzein et al. (2017) for an alternative interpretation of such findings.)

So whereas the feelings of agency and familiarity are both consequences of unexpected fluency of processing, the feeling of surprise is supposed to be the opposite: it is a consequence of unexpected disfluency.[2]

1.6. Metacognitive Feelings as Sensations

Metacognitive feelings are aspects of the overall phenomenal character of experiences which their subjects take to be informative about things that are only distantly related (if at all) to the things that those experiences intentionally relate the subject to.[3]

To illustrate, having a feeling of familiarity is not a matter of standing in any intentional relation to the property of familiarity, but it is something that we can interpret as informative about familiarity.[4]

We might think of metacognitive feelings as lacking intentional objects altogether; this would make them like sensations in Reid (1785)'s sense. Not everyone accepts that such things could exist, of course (because they aim to explicate phenomenology in terms of intentional content or whatever). We can be agnostic by noting that nothing is lost by treating metacognitive feelings as if they were sensations.

Sensations are:

---

[2] An alternative is proposed by Foster & Keane (2015, p. 79): 'the MEB theory of surprise posits that: Experienced surprise is a metacognitive assessment of the cognitive work carried out to explain an outcome. Very surprising events are those that are difficult to explain, while less surprising events are those which are easier to explain.' Foster & Keane (2015, p. 79) is about reactions to reading about something unexpected, whereas Reisenzein (2000) measures how people experience unexpected events (changes to stimuli while solving a problem). The latter is much closer to our concerns. But the truth of either account of surprise, or of an account combining the two insights, would indicate that there is a metacognitive feeling of surprise.

[3] This is consistent with, but weaker than, Koriat's theory: 'metacognitive feelings are mediated by the implicit application of nonanalytic heuristics … [which] operate below full consciousness, relying on a variety of cues … [and] affect metacognitive judgments by influencing subjective experience itself' (Koriat 2000, p. 158; see also Koriat 2007, pp. 313–5).

[4] Why accept this? You cannot perceive familiarity or agency any more than you can perceive electricity. Perceptual processes do not reach far back into your past, nor are they concerned with questions about whether you are the agent of an action. So to think that metacognitive feelings intentionally relate you to facts about familiarity or agency requires postulating a novel kind of sensory process, some kind of inner or bodily sense.

- monadic properties of events, specifically perceptual experiences,
- individuated by their normal causes—in the case of feelings of familiarity, its normal cause is ease of processing
- which alter the overall phenomenal character of those experiences
- in ways not determined by the experiences' contents (so two experiences can have the same content while one has a sensational property which the other lacks).

If this is right, why do metacognitive feelings invite judgements? Why does the feeling of familiarity (say) even so much as nudge you to judge that the face photographed here is familiar to you? (This is roughly Dokic (2012)'s question.)

1.7. How Metacognitive Feelings Link Fast to Slow Processes

The feeling of familiarity is reliably caused by things which are familiar. This is because in a limited, but useful, range of cases, things which you can process fluently are things which are familiar to you. After all, familiarity is one (of several) causes of ease of processing.

Over time you learn, perhaps implicitly, to associate the feeling of familiarity with things being familiar. (Although you can unlearn this association in a carefully controlled experimental setting; Wan et al. 2008.)

So a fast processes causes a feeling, which triggers a learned association, which in turn biases a slow process to determine that the likely cause of the feeling is familiar.

Could this be a model for how fast processes influence slow processes generally?

2. Conclusion: Six Questions

We started by asking six questions. In conclusion, we review discoveries about each.

2.1. In Which Domains Is There Substantial Evidence for a Two Systems Theory?

Substantial evidence is evidence from multiple studies from different labs using different approaches.

We have seen that there is substantial evidence for two systems theories of mindreading (in *Mindreading: Signature Limits, and Development* in Lecture 02) and ethics (in *Ethical Cognition* in Lecture 03). We have also seen some evidence for two systems theories of physical cognition (in *Speed-Accuracy Trade-Offs (in Physical Cognition)* in Lecture 01). In no case is the evidence sufficient to entirely rule out alternative, one system theories.

There are also many other cases we did not consider (as listed in *The Core Idea* in Lecture 01). In some of these cases a two systems theory is well established (e.g. memory, number and instrumental behaviour).

2.2. How Are the Two Systems Distinguished?

Our approach was to consider a stripped-down, core claim about processes which differ in how fast they are. Further respects in which systems can be distinguished—by appeal to automaticity, say—are captured by auxiliary hypotheses (see *The Core Idea* in Lecture 01).

We saw that two systems for mindreading can be distinguished by (i) the different degrees to which they are automatic (see *Mindreading: Automaticity* in Lecture 02) and (ii) the different models of minds and actions they employ (see *Mindreading: Signature Limits, and Development* in Lecture 02).

Two systems for physical cognition can be distinguished by the range of models of the physical which can characterise their operations. One system appears limited to impetus mechanics while the other is more flexible. The more limited system also appears to be partly responsible for representational momentum and perhaps other broadly perceptual effects, and thereby to influence the overall phenomenal character of experiences (see *Speed-Accuracy Trade-Offs (in Physical Cognition)* in Lecture 01).

In the ethical domain, two systems theories are wildly accepted but there is much uncertainty about what distinguishes them. In our view this is a significant open challenge (see *Ethical Cognition* in Lecture 03).

2.3. What, If Any, Kind of Unity Is There Across Domains?

We did not identify substantial evidence for a hypothesis which generates readily testable predictions and could be used to characterise features of two systems in different domains.

Features commonly conjectured as common themes across domains include automaticity, informational encapsulation and domain specificity. We observed that there is not a lot of evidence to support these conjectures. For instance, evidence for or against automaticity is hard to identify in the domains of ethical and physical cognition.

2.4. Why Are There Two Systems?

- speed–accuracy trade-offs: different challenges call for different trade-offs between speed and accuracy. Having more than one system allows for radically different trade-offs between speed and accuracy. (This was illustrated in *The Core Idea* in Lecture 01.)
- learning and development: having more than one system where the fast system is relatively unchanging over development can provide an optimal balance between reliably meeting everyday practical needs and making it possible to pursue learning where there is a high risk of error but also a large potential reward. (This was illustrated in *Mindreading: Signature Limits, and Development* in Lecture 02.)
- phylogeny and culture: the historical emergence of writing was a consequence of a slow system building on abilities made possible by some fast systems. As this suggests, there are some things best provided phylogenetically (or at least through learning processes that do not depend on large-scale cooperative cultural projects) and others that can be provided through large-scale cooperative cultural projects.

2.5. When, If Ever, Are Two Systems Better Than One?

If you are building a survival system you want quick and dirty heuristics that are good enough to keep it alive: you don't necessarily care about the truth. If, by contrast, you are building a thinker, you want her to be able to think things that are true irrespective of their survival value. This cuts two ways. On the one hand, you want the thinker's thoughts not to be constrained by heuristics that ensure her survival. On the other hand, in allowing the thinker freedom to pursue the truth there is an excellent chance she will end up profoundly mistaken or deeply confused about the nature of physical objects. If she turns to philosophy, she may even end up convincing herself that nothing exists apart from her. So you don't want thought contaminated by survival heuristics and you don't want survival heuristics contaminated by thought. Or if some contamination is inevitable, you at least want to limit it. This is beautifully achieved by giving your thinker two (or more) systems, one fast and the other slow. Providing, of course, that the two are not directly connected but rather linked only very loosely, via intentional isolators like metacognitive feelings.

2.6. How, If At All, Do the Two Systems Interact? What Are the Barriers to Interaction Between Them?

Because of how we characterised what it is for systems to be distinct, there is a tension between postulating two (or more) systems and postulating interactions between them. We suggested that the distinctness of systems consists in there being processes which differ in conditions which influence whether they occur, and which outputs they generate (in *The Core Idea* in Lecture 01). As the scope for interaction increases, the grounds for distinguishing systems weaken.

In both mindreading and physical cognition, we saw that it is possible for distinct processes to yield incompatible outputs in response to a single stimulus. Importantly, in the case of mindreading we also saw that this can work both ways: there are situations in which fast processes support correct responses while slow processes support incorrect responses; and conversely (see *Mindreading: Signature Limits, and Development* in Lecture 02). This suggests that there are barriers to interaction between systems. And perhaps that the representations they operate over are not inferentially integrated.

One conjecture, which we did not explore in depth, is that fast and slow processes differ in operating over representations which differ in format. This barrier to interaction may explain the lack of inferential integration.

We saw that it is possible for a fast process to influence a slow one indirectly and asynchronously if the fast system can modify the overall phenomenal character of experiences (see *Speed-Accuracy Trade-Offs (in Physical Cognition)* in Lecture 01). This provides one model for understanding interactions between fast and slow systems.

It is also possible that metacognitive feelings provide a way for fast processes to influence slow processes synchronously (see *Metacognitive Feelings: How Do Fast and Slow Processes Interact?* (section §1)).

Glossary

automatic On this course, a process is *automatic* just if whether or not it occurs is to a significant extent independent of your current task, motivations and intentions. To say that *mindreading is automatic* is to say that it involves only automatic processes. The term 'automatic' has been used in a variety of ways by other authors: see Moors (2014, p. 22) for a one-page overview, Moors & De Houwer (2006) for a detailed theoretical review, or Bargh (1992) for a classic and very readable introduction 7, 10

cognitively efficient A process is *cognitively efficient* to the degree that it does not consume working memory and other scarce cognitive resources. 10

domain specific A process is domain specific to the extent that there are limits on the range of functions its outputs typically serve. Domain-specific processes are commonly contrasted with general-purpose processes. 7

fast A *fast* process is one that is to to some interesting degree cognitively efficient (and therefore likely also some interesting degree automatic). These processes are also sometimes characterised as able to yield rapid responses.

Since automaticity and cognitive efficiency are matters of degree, it is only strictly correct to identify some processes as faster than others.

The fast-slow distinction has been variously characterised in ways that do not entirely overlap (even individual author have offered differing characterisations at different times; e.g. Kahneman 2013; Morewedge & Kahneman 2010; Kahneman & Klein 2009; Kahneman 2002): as its advocates stress, it is a rough-and-ready tool rather than an element in a rigorous theory. 2, 6–9, 11

inferential integration For states to be *inferentially integrated* means that: (a) they can come to be nonaccidentally related in ways that are approximately rational thanks to processes of inference and practical reasoning; and (b) in the absence of obstacles such as time pressure, distraction, motivations to be irrational, self-deception or exhaustion, approximately rational harmony will characteristically be maintained among those states that are currently active. 2, 9

informational encapsulation One process is informationally encapsulated from some other processes to the extent that there are limits on the one process' ability to consume information available to the other processes. (See Fodor 1983; Clarke 2020, pp. 5ff.) 7

intentional isolator An event or state which links representations but either lacks intentional features entirely or else has intentional features that are only very distantly related to those of the two representations it links. Metacognitive feelings and behaviours are paradigm intentional isolators. 8

metacognitive feeling A metacognitive feeling is a feeling which is caused by a metacognitive process. Paradigm examples of metacognitive feel-

ings include the feeling of familiarity, the feeling that something is on the tip of your tongue, the feeling of confidence and the feeling that someone's eyes are boring into your back. On this course, we assume that one characteristic of metacogntive feelings is that either they lack intentional objects altogether, or else what their subjects take them to be about is typically only very distantly related to their intentional objects. (This is controversial—see Dokic 2012 for a variety of conflicting theories.) 2, 4, 8–10

metacognitive process A process which monitors another cognitive process. For instance, a process which monitors the fluency of recall, or of action selection, is a metacognitive process. 3, 10

representational format Format is an aspect of representation distinct from content (and from vehicle). Consider that a line on a map and a list of verbal instructions can both represent the same route through a city. They differ in format: one is cartographic, the other linguistic. 9

slow converse of fast. 2, 6, 8, 9

References

Bargh, J. A. (1992). The Ecology of Automaticity: Toward Establishing the Conditions Needed to Produce Automatic Processing Effects. *The American Journal of Psychology*, *105*(2), 181–199.

Brown, A. S. (1991). A review of the tip-of-the-tongue experience. *Psychological Bulletin*, *109*(2), 204–223.

Brown, A. S. & Marsh, E. J. (2010). Digging into Déjà Vu: Recent Research on Possible Mechanisms. In B. H. Ross (Ed.), *Psychology of Learning and Motivation*, volume 53 of *The Psychology of Learning and Motivation: Advances in Research and Theory* (pp. 33–62). Academic Press.

Clarke, S. (2020). Cognitive penetration and informational encapsulation: Have we been failing the module? *Philosophical Studies*.

Dokic, J. (2012). Seeds of self-knowledge: noetic feelings and metacognition. In M. J. Beran, J. L. Brandl, J. Perner, & J. Proust (Eds.), *Foundations of metacognition* (pp. 302–321). Oxford University Press Oxford, England.

Fodor, J. (1983). *The Modularity of Mind: an Essay on Faculty Psychology*. Bradford book. Cambridge, Mass ; London: MIT Press.

Foster, M. I. & Keane, M. T. (2015). Why some surprises are more surprising than others: Surprise as a metacognitive sense of explanatory difficulty. *Cognitive Psychology*, *81*, 74–116.

Haggard, P. & Chambon, V. (2012). Sense of agency. *Current Biology*, *22*(10), R390–R392.

Kahneman, D. (2002). Maps of bounded rationality: A perspective on intuitive judgment and choice. In T. Frangsmyr (Ed.), *Le Prix Nobel, ed. T. Frangsmyr, 416–499.*, volume 8 (pp. 351–401). Stockholm, Sweden: Nobel Foundation.

Kahneman, D. (2013). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kahneman, D. & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*(6), 515–526.

Koriat, A. (2000). The Feeling of Knowing: Some Metatheoretical Implications for Consciousness and Control. *Consciousness and Cognition*, *9*(2), 149–171.

Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge Handbook of Consciousness* (pp. 289–325). New York, NY, US: Cambridge University Press.

Moors, A. (2014). Examining the mapping problem in dual process models. In *Dual process theories of the social mind* (pp. 20–34). Guilford.

Moors, A. & De Houwer, J. (2006). Automaticity: A Theoretical and Conceptual Analysis. *Psychological Bulletin*, *132*(2), 297–326.

Morewedge, C. K. & Kahneman, D. (2010). Associative processes in intuitive judgment. *Trends in Cognitive Sciences*, *14*(10), 435–440.

Reid, T. (1785). *An Inquiry into the Human Mind* (Fourth Edition ed.). London: T. Cadell et al.

Reisenzein, R. (2000). The subjective experience of surprise. In H. Bless & J. P. Forgas (Eds.), *The message within: The role of subjective experience in social cognition and behavior* (pp. 262–279). Hove: Psychology Press.

Reisenzein, R., Horstmann, G., & Schützwohl, A. (2017). The Cognitive-Evolutionary Model of Surprise: A Review of the Evidence. *Topics in Cognitive Science*, n/a–n/a.

Scott, R. B. & Dienes, Z. (2008). The conscious, the unconscious, and familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(5), 1264–1288.

Sidarus, N., Chambon, V., & Haggard, P. (2013). Priming of actions increases sense of control over unexpected outcomes. *Consciousness and Cognition*, *22*(4), 1403–1411.

Sidarus, N., Vuorre, M., & Haggard, P. (2017). How action selection influences the sense of agency: An ERP study. *NeuroImage*, *150*, 1–13.

Velasco, P. F. & Casati, R. (2020). Subjective disorientation as a metacognitive feeling. *Spatial Cognition & Computation*, *20*(4), 281–305.

Wan, L., Dienes, Z., & Fu, X. (2008). Intentional control based on familiarity in artificial grammar learning. *Consciousness and Cognition*, *17*(4), 1209–1218.

Wenke, D., Fleming, S. M., & Haggard, P. (2010). Subliminal priming of actions influences sense of control over effects of action. *Cognition*, *115*(1), 26–38.

Whittlesea, B. W. A. & Williams, L. D. (1998). Why do strangers feel familiar, but friends don't? a discrepancy-attribution account of feelings of familiarity. *Acta Psychologica*, *98*(2-3), 141–165.

Widner, R. L., Otani, H., & Winkelman, S. E. (2005). Tip-of-the-Tongue Experiences Are Not Merely. *The Journal of General Psychology*, *132*(4), 392–407.