Synthese (2021) 198:6573–6596
https://doi.org/10.1007/s11229-019-02477-2

FOLK PSYCHOLOGY: PLURALISTIC APPROACHES

A new perspective on the relationship between metacognition and social cognition: metacognitive concepts as socio-cognitive tools

Tadeusz W. Zawidzki[1]

Received: 17 May 2019 / Accepted: 11 November 2019 / Published online: 15 November 2019

Abstract

I defend an alternative to the two traditional accounts of the relationship between metacognition and social cognition: metacognition as primary versus social cognition as primary. These accounts have complementary explanatory vices and virtues. They also share a natural assumption: that interpretation in terms of mental states is "spectatorial", aiming exclusively for an objective description of the mental facts about self and others. I argue that if one rejects this assumption in favor of the view that interpretation in terms of mental states also plays important regulative roles with respect to minds and behavior, a new and superior conception of the relationship between metacognition and social cognition comes into view. On this conception, person-level metacognitive concepts are socio-cognitive tools that shape us into better cognitive agents and more predictable cognitive objects, thereby enhancing our abilities at social coordination. Mastery of these metacognitive concepts relies on subpersonal, non-conceptual, procedural metacognition. This reconceptualization of the relationship between metacognition and social cognition combines the complementary explanatory virtues of the two traditional conceptions, while avoiding their complementary explanatory vices.

Keywords Social cognition · Metacognition · Metacognitive concepts · Procedural metacognition · Mental state attribution · Folk psychology as regulative · Socio-cognitive tools

✉ Tadeusz W. Zawidzki
zawidzki@gwu.edu

[1] Department of Philosophy, George Washington University, Washington, DC 20052, USA

1 Introduction

Metacognition is cognition about cognition, while social cognition is cognition about social phenomena, primarily constituted by the behavior of a cognitive agent's fellow cognitive agents, typically conspecifics with whom she regularly interacts. Although these are different domains, philosophers and cognitive scientists have overwhelmingly assumed that they are intimately related. The reason is straightforward: the behavior that constitutes the social domain is largely caused by the cognitive states of the agents who engage in this behavior; hence, it is reasonable to expect cognition about the social domain to depend upon cognition about cognition.

Metacognition, however, has important non-social applications as well. The reason is that the cognitive agent who thinks about the cognitive states of her fellows also instantiates her own cognitive states. Hence, cognitive agents can think about their own cognitive states: the central non-social application of metacognition. Therefore, it is unsurprising that philosophy and cognitive science have devoted a lot of theoretical effort to understanding the relationship between the non-social, i.e., self-directed, and social applications of metacognition. This centuries-old enterprise has yielded two rival conceptualizations of the relationship between metacognition and social cognition, which I shall call, respectively, the metacognition-as-primary view (henceforth "MP") and the social-cognition-as-primary view (henceforth "SP").

According to MP, the application of metacognition to the social domain is secondary, and derives from prior, non-social applications to the self. The classic development of this idea is Descartes' argument in the *Meditations* (1641/1993), according to which one first introspectively establishes the contents of one's own mind, before inferring that other human bodies one observes behaving similarly to one's own are animated by minds with similar contents. This argument from analogy was later adopted and adapted by other philosophers for similar purposes, e.g., Hume's argument concerning non-human minds (1739/1978) and Russell's argument for believing in other minds (1948/2009). It also forms the basis for one of the two chief rival accounts of social cognition in contemporary cognitive science: the simulation theory, according to which we learn the mental states of our fellows by pretending to be in their situations, noting the mental states we token as a result, and projecting these onto our interpretive targets (Gordon 1986; Goldman 2006).

According to SP, the non-social application of metacognition to oneself is secondary, and derives from prior, social applications to others. Here, by "prior", I do not necessarily intend temporal priority. Rather, the point is that, according to SP, the primary function of metacognition is social, i.e., understanding the domain of publicly observable behavior by cognitive agents. Since the agent who gains such understanding is herself a cognitive agent generating publicly observable behavior, the social understanding gained via metacognition subsumes the subject's own behavior as a special case. So, even if self and other understanding arise simultaneously, the former derives from the latter since it constitutes merely a special case of the latter.

SP has a more recent provenance than MP, tracing to Wilfrid Sellars's thought experiments in his "Empiricism and the Philosophy of Mind" (1956). According to Sellars, metacognition begins as a scientific hypothesis about the etiology of publicly observable, intelligent behavior, by self or others, later taking on a non-inferential, "reporting" role applied to the self, in the way that theoretical concepts in science can take on non-inferential reporting roles when scientists are trained to directly recognize the presence of theoretical entities, without having to explicitly infer them from observable evidence. This perspective has been taken up by philosophers and cognitive scientists in defense of the chief contemporary rival account of social cognition to simulation theory: the theory–theory (Wellman 1990; Gopnik and Meltzoff 1996). According to the theory–theory, children acquire metacognitive capacities when they formulate hypotheses about the hidden etiologies of publicly observable, intelligent behavior, both of others and their own. These hypotheses are formulated in terms of concepts of mental states, and are then verified via behavioral tests. We thus learn about others' mental states and our own roughly simultaneously, as part of the project of constructing empirically verified theories of publicly observable, intelligent behavior.

The theoretical virtues and vices of MP and SP are, roughly, mirror images of each other: each strength of the former constitutes a weakness of the latter, and vice versa. For this reason, it is unsurprising that most recent models of social cognition incorporate both into hybrid accounts. In addition, both MP and SP share a more fundamental assumption, which, following Hutto (2008), I call the "spectatorial" conception (henceforth, "SC") of social cognition, metacognition, and folk psychology more broadly. According to SC, the point of our interpretive practices, whether self-directed, as in non-social metacognition, or other-directed, as in social cognition, is to *describe*, or construct an accurate picture of the relevant domain, whether it be one's own mind, or the behavior and minds of others. MP, SP, and their various hybrids are all versions of SC. The goal of this paper is to explore an alternative conception of the relations between metacognition and social cognition that results from abandoning SC: how we might conceive of the relationship between metacognition and social cognition if we assume they aim *not only* to accurately describe the minds and behavior of oneself and others but, also, to *regulate* the minds and behavior of oneself and others. I shall argue that the reconceptualization I outline avoids the theoretical vices of, and embodies the theoretical virtues of both MP and SP.

The discussion proceeds as follows. Section 2 details the relative explanatory advantages and disadvantages of MP and SP. Section 3 addresses the motivations for abandoning SC, and provides details about the regulative alternative to it. Section 4 draws on Proust's (2013) distinction between procedural and conceptual metacognition to formulate a new conception of the relationship between metacognition and social cognition suitable for the regulative alternative to SC. Section 5 argues that this new conception of the relationship between metacognition and social cognition avoids the theoretical vices and embodies the theoretical virtues of both MP and SP.

2 The pros and cons of metacognition as primary versus social cognition as primary

As I mentioned in the introductory section, MP and SP have complementary explanatory vices and virtues. These are widely recognized, and motivate many to hold out hope for hybrid models of human social cognition, in which the two conceptions compensate for each other's shortcomings. Here, I recount these complementary vices and virtues in greater detail.

The most obvious benefit of taking self-directed metacognition as primary is the intuitive or phenomenological plausibility of this assumption. This explains why Descartes' arguments in the *Meditations* are so seductive. It is hard to question the assumption that we have much more direct and reliable access to our own cognitive states than to those of others. Furthermore, it is plausible that we do not really understand others unless we are able to experience the world from their perspective, to walk in their shoes, so to speak. This appears to be the commonsense understanding of empathy, as expressed in many of the world's religious traditions, for example. And the most natural way to explain this conception of empathy is in terms of MP: we have primitive and direct knowledge of what it is like to experience the world from our own perspective, and we empathize with others when we pretend to be in their situations and project our resulting first-person experiences onto them.

Another, closely related benefit of taking self-directed metacognition as primary is that it makes sense of the manifest asymmetries between self-directed metacognition and social cognition. Our access to our own cognitive states seems to be direct and epistemically authoritative in a way that our access to the cognitive states of others is not. We do not *appear* to rely on behavioral evidence to determine our own cognitive states; yet, this is unavoidable when it comes to the cognitive states of others. Furthermore, while it is typical for us to be wrong and corrigible about the cognitive states of others, it is *not* typical, at least in everyday practice, for us to be wrong or corrigible about our own cognitive states. If self-directed metacognition is primary and social cognition derives from it then these asymmetries are unsurprising: our knowledge of others' cognitive states is based on a fraught projection from our far more secure knowledge of our own cognitive states. However, if social cognition is primary, and self-directed metacognition a mere special case of it, then these asymmetries in epistemic access and authority are *prima facie* mysterious.

Finally, defenders of the simulation theory propose the following empirical argument in favor of it and against the theory–theory (Gordon 1986; Goldman 1989). The task of interpreting the behavior of others via simulation is computationally more tractable than the task of interpreting the behavior of others via theory. If there are laws linking mental states to each other and behavior, as the theory–theory supposes, then these laws are bound to be extremely complex, and hedged with multifarious "ceteris paribus" clauses (Gauker 2003, p. 240). The idea that quotidian interpretation requires computations over such complex laws seems implausible: typically, such interpretation is quick and seamless, and computational tractability issues do not tend to arise. But, according to simulation theory, interpreters needn't even *know* any laws, let alone compute over them, in order to interpret others. Instead, interpreters unconsciously use their own cognitive economy as a model of those of others. When we interpret others, we pretend to be in their situations, and then simply let our own minds react as they would, were we really in those situations. No computation over complex laws is necessary, hence issues of computational tractability do not arise. Thus, simulation theory makes better sense of the evident fluidity of quotidian interpretation than theory–theory. Since the most prominent versions of the simulation theory are based on MP, while the most prominent versions of the theory–theory are based on SP, this empirical argument in favor of simulation theory constitutes an explanatory advantage of MP over SP.

Thus, MP has the following three explanatory advantages over SP: (1) it is intuitively more plausible; (2) it makes sense of asymmetries in epistemic access and authority between metacognition and social cognition; (3) it is less susceptible to issues of computational tractability. However, SP has four explanatory advantages of its own over MP. First, the direct epistemic relation that agents have to their own cognitive states, posited by MP, is very mysterious. How is cognitive science to make sense of such direct access, unmediated by any form of behavioral evidence? Perhaps such self-transparency makes sense for a Cartesian dualist, but if metacognition is susceptible to mechanistic explanation, then there has to be some physically specifiable means by which agents gain epistemic access to their own cognitive states. SP at least has a clear proposal regarding this: as in the case of attributing mental states to others, attributing mental states to oneself involves inferring the presence of mental states based on behavioral evidence (Carruthers 2011).

The second explanatory advantage of SP over MP concerns the reliability of social cognition. According to MP, we make sense of others by projecting our own mental states onto them. But why suppose that such a procedure would yield reliable interpretations of others? After all, people vary dramatically in background assumptions and preferences regarding the world. If I simply take for granted that others assume and are motivated by the same things as myself, then I am bound to misinterpret others a lot of the time (Weiskopf 2005). But, typically, we do not seem to be so unreliable in our interpretations of others. Indeed, it is hard to explain how our interpretive practices could persist in the face of such unreliability. SP at least has the beginnings of an explanation of why our interpretations tend to be reliable: they are products of a science-like theory of publicly observable, intelligent behavior, that has undergone empirical test and confirmation through phylogeny and ontogeny.

A third explanatory advantage of SP over MP concerns the concepts deployed in metacognition and social cognition. MP has a harder time explaining why we would use the *same* concepts to interpret ourselves and others. These concepts are concepts of mental states, like beliefs, desires, intentions, and emotions. However, consider how different are the conditions and consequences of applying these concepts to oneself as compared to others. According to MP, we learn about our own mental states through a mysteriously private process of introspection, but we learn about others' mental states through inferences from observations of publicly accessible behavior. Furthermore, we represent our own mental states in ways that yield a kind of direct control unavailable in the case of the mental states of others. Representing an unwanted desire to smoke as my own immediately gives rise to attempts to curb its influence, but such attempts to curb unwanted desires to smoke in others are unavailable to me. Given these different profiles in the conditions and consequences of applying mental state concepts to oneself and to others, it is puzzling that we use the same concepts to interpret ourselves and others. Why would we deploy the same concepts for such different jobs? SP has an answer to this question. Metacognition is a special case of social cognition. The concepts we use to interpret others and ourselves are products of the same science-like process of hypothesis and test applied to publicly observable behavior. We may later discover that there are unique forms of evidence and practical import when the behavior we interpret is our own, but the origins of the concepts used to interpret ourselves and others is the same, i.e., explaining publicly observable behavior; hence, it is no surprise that the concepts employed are the same.

The final explanatory advantage of SP over MP concerns some noteworthy empirical evidence from developmental psychology. It is now well established that children learn to apply mental state concepts to themselves and others at the same time. For example, children younger than 4 years of age, who mistakenly assume that others have true beliefs about situations of which they know them to be ignorant, also mistakenly assume that they themselves had true beliefs about situations of which they know themselves to have been ignorant. Children stop making this mistake, about themselves and others, at round 4 years of age (Gopnik and Astington 1988). This is widely taken to show that children learn to apply the concept of false belief to themselves and others at the same time, as SP predicts. However, this result is hard to explain for MP. If social cognition derives from prior metacognition, we would expect children to correctly apply mental state concepts to themselves before they correctly applied them to others.

Thus, the three advantages of MP over SP are countered by four advantages of SP over MP: (1) SP avoids appeals to mysterious processes of introspection; (2) SP appears better poised to explain the reliability of social cognition; (3) SP provides some explanation about why we use the same concepts in metacognition as in social cognition; (4) SP makes sense of some prominent empirical results regarding the ontogeny of social cognition.

This explanatory impasse has motivated a growing consensus on hybrid models of social cognition (Nichols and Stich 2003; Goldman 2006; Carruthers 1996). The hope is that the strengths of each perspective can compensate for the weaknesses of the other. However, it is unclear how this is supposed to work in detail. For example, suppose a model proposes supplementing a version of MP, like simulation theory, with a version of SP, like theory–theory, to address the problem of reliability: we can reliably interpret others who are different from us because theory-like processes take over from simulation in such cases (Nichols and Stich 2003, p. 140; Goldman 2006, p. 184). But then how do we avoid issues of computational tractability? The problem is that these strengths and weaknesses of MP and SP are *complementary*. SP explains reliability better than MP, *at the cost of failing to explain computational tractability*, and vice versa. The same can be said about the other relative trade-offs: intuitive plausibility versus less mystery, explaining asymmetries versus explaining use of the same concepts, explaining some empirical results versus explaining others [while theory–theory does better at explaining the onset of mental state concepts in ontogeny, simulation theory arguably does better at explaining well-known egocentric biases, like the "curse of knowledge" (Goldman 2006, p. 168), than theory–theory (Nichols and Stich 2003, pp. 70–73, 80–81; Goldman 2006, p. 173)].

I do not claim that any of these problems are devastating for either SP or MP. Defenders of both perspectives are well aware of the problems, and propose ingenious solutions. However, consensus remains elusive, despite the recent turn to hybrid models. Given this impasse, it is at least worth exploring alternatives to the MP versus SP dialectic, should any be available. In what follows, I argue that, once we jettison the assumption, common to SP and MP, that interpretation aims primarily to *describe independent facts about mental states*, and fully appreciate its significant and diverse *regulative* impacts on mental states and the behavior they cause, it is possible to reconceive the relations between metacognition and social cognition in a way that neatly avoids the problems with SP and MP discussed above. I now turn to a discussion of what more interpretation might aim at, if not only true descriptions of independent facts about mental states.

3 Against the spectatorial conception: interpretation as regulative practice

As I noted in the introductory section, both MP and SP make a very natural assumption about the aim of interpretation: mental state attribution aims to construct accurate, objective descriptions of facts about mental states. Even in the case of self-interpretation, though the domain being described, i.e. the subject's own mental states, is subjective, the goal of metacognition is an accurate, shareable description of the facts about that domain; so, in this sense, non-social metacognition aims at an objective description, even if it is destined to fail due to some residual subjective aspects that it cannot capture. As Hutto puts it, regarding the interpretation of others:

> [P]hilosophers have tended to make a number of ultimately questionable interrelated assumptions about the context in which we engage in commonsense psychology, assumptions that affect our thinking about its very nature. Chief amongst these is that we are normally at a theoretical remove from others. The attitude we adopt towards others is thus on a par with that deployed when understanding 'foreign bodies' quite generally: We ascribe causally efficacious inner mental states to them for the purpose of prediction, explanation and control. As a consequence, this fosters the idea that our initial stance with respect to others is essentially estranged. (2004, p. 549)

McGeer makes a similar point about self-interpretation:

> [There is a] fiction to which many philosophers and psychologists have clung … that self-knowledge consists in coming to know (perhaps via theoretical mediation) a collection of facts: facts about one's own thoughts, feelings,

> intentions, and so on … the reporter-predictor model of authoritative self-knowledge… (1996, p. 506)

But, as Hutto, McGeer, and others (Morton 2003; Zawidzki 2008) have persuasively argued, there are good reasons to question this assumption. The key problem is noted by Hutto: it is misleading to conceive of the relations between interpreters and their targets on the model of how we understand "foreign bodies" more generally; our initial stance with respect to others is *not* so essentially estranged. McGeer identifies one important reason for this: we do *not* "interact with one another as scientist to object, as observer to observed, since the 'objects' themselves—viz. … [our]selves and other agents—are changing under pressure of the 'explanatory-predictive' attributions that are made to them" (2007, p. 146). This is most obvious in the case of self-interpretation, as McGeer notes elsewhere:

> we are able to ensure a fit between the psychological profile we create of ourselves in first-person utterances and the acts our self-attributed intentional states are meant to predict and explain simply by adjusting our actions in appropriate ways. Thus, because we do not just wait to see if our actions make sense in light of intentional self-attributions, but rather *make* them make sense, the tale we tell of ourselves from the intentional stance is importantly unlike the tale we tell of other people (or even of other things). I cannot make it the case that you behave in ways coherent with what I say you hope, desire, or fear any more than I can make it the case that the world is a certain way by announcing how (I think) it is; but I can and do govern my own actions in ways that fit with the claims I make about myself. (1996, p. 507)

Furthermore, as McGeer acknowledges in other work (2001, 2007, 2015), there are also less direct means by which our interpretations of others can exert regulative pressures on them.

Most obviously, caretaker interpretations of their infant charges can come to regulate infant behavior (McGeer 2001). Mameli (2001) speculates that "social expectancies" might turn interpretations into self-fulfilling prophecies, as when gendered, adult interpretations of neutral infant behavior lead infants to conform to the adult social expectations that accompany such interpretations. For example, a fussy infant might be interpreted as distressed if perceived as female but angry if perceived as male, with all the differences in social expectancies that accompany attributions of distress and anger (Golombok and Fivush 1994). The infant then learns to conform to expectations that typically accompany attributions of anger if male, and distress if female, even if the initial behavior that triggered the adult interpretations was neutral between them. Mameli (2001) also suggests that such mechanisms might boot-strap the capacity for intentional communication in infancy, when caretakers interpret initially unintentional acts, like certain arm motions, as intentional, e.g., as points or requests. There is also evidence that children internalize interpretations of adult models, which they then use to self-regulate via self-interpretation. For example, children socialized to believe that academic success is caused by innate, immutable talent tend to be less flexible and resilient in response to failures in academic tasks, and, consequently, less likely to learn challenging material, thereby turning the initial self-interpretation into a self-fulfilling prophecy (Nix et al. 2015). Finally, Hacking (1995) argues that "looping effects" typify "human kinds", such as those used in psychiatric classification and other areas of the social sciences, distinguishing them from natural kinds precisely in virtue of the regulative and other effects they have on the domain they purport to describe. For example, Nineteenth Century psychiatric patients diagnosed with "fugue" syndrome tended to actively shape their own behavior such that it conformed to the widely, publicly disseminated symptomology that characterized this spurious condition (Hacking 1998).

Thus, there are good reasons to suppose that interpretations of self and other often play regulative roles with respect to behavior. But why think this is an important function of interpretation? Such regulative effects might be relatively marginal, compared to the descriptive function typically attributed to our interpretive practices. There are a number of persuasive reasons to resist this very natural view, i.e., reasons why the regulative function of interpretation is at least as important as its descriptive function, and, in some cases at least, enables the latter. Before discussing these reasons, it is important to clarify exactly what I mean by "interpretation", for the purpose of this discussion. My focus is on person-level, linguistically expressible, reflective interpretation of behavior in terms of high-level, meta-cognitive concepts, like the propositional attitudes and emotions that have lexicalized expressions. This is an important qualification, because a number of theorists have proposed "lower-level" forms of interpretation that make possible various forms of social intelligence in non-human animals and human infants. For example, Butterfill and Apperly (2013) appeal to a "minimal theory of mind" in order to explain successful performance of human infants on non-verbal variants of the "false belief task". They argue that such behavior can be explained in terms of a kind of interpretation that involves sub-personal attribution of non-propositional, relational states, like goals, perceptions, and encodings, directed at object locations. Moore (2017, 2018) argues that similarly low-level forms of interpretation can explain Gricean forms of communication in non-humans, and in human infants acquiring a first language.[1] What I say about the regulative functions of interpretation below is *not* intended to apply to these kinds of interpretation. In fact, it is likely that the kinds of regulative functions I argue "high-level" interpretation performs are made possible by the descriptive and predictive functions of these "lower-level" forms of interpretation. The classic versions of MP and SP, in both philosophy and cognitive science, focus, almost exclusively, on the interpretation of behavior in terms of person-level, linguistically expressible, reflective meta-cognitive concepts, like the propositional attitudes and lexicalized emotion categories. MP and SP both assume SC regarding this form of interpretation, and this assumption is the target of my arguments in what follows.

The first reason why the regulative functions of interpretation should not be dismissed as marginal is that mental state attributions often appear to display what Searle (1979) calls the "world-to-word" direction of fit typical of imperative speech acts rather than descriptive speech acts like assertions, which display

[1] Although Moore does not think of this form of interpretation as "sub-personal", but, rather, as an "unreflective and undemanding personal level" phenomenon. (personal communication).

a "word-to-world" direction of fit. If the world does not match the content of an assertion, as in a false description like "The sun is shining" uttered at night, there is normative pressure to alter the assertion such that it better describes the world. Conversely, if the world does not match the content of an imperative, as in an unobeyed command, like a soldier failing to come to attention in response to a superior's "Attention!", there is normative pressure to alter the world, i.e., the soldier's behavior, such that it better matches the content of the command. Mental state attributions and other interpretive acts often seem to instantiate the latter direction of fit rather than the former. For example, if someone says "Team A will definitely beat Team B", triggering an attribution of the belief that Team A will definitely beat Team B, but then proceeds to bet an enormous amount of money on Team B, the belief attribution *can* be withdrawn and replaced, as for a description, *but there is always another option*: to criticize the bet as inconsistent with the belief attribution, in an attempt to alter that behavior, as for a command.[2] If interpretation were like our other descriptive practices, then this pattern would be puzzling. For most of our descriptive practices, holding onto claims at odds with the world they purport to describe and attempting to alter the world such that it matches their content is *never* an option. When one says it is sunny on a cloudy day, one never has the option to maintain the assertion and try to change the weather. One reason for this disanalogy between interpretation and other descriptive practices is the intimate link between assumptions of rationality and interpretation (Davidson 1984; Dennett 1987). If interpretations typically involve the assumption that the interpretive target is rational, then it is unsurprising that they appear to sometimes display a world-to-word direction of fit.

A second reason to doubt that interpretation aims primarily at accurate description is that it appears very poorly designed for this task. The reason is holism: the relations between mental states and their observable triggers and consequences are many–many, highly complex, and hedged with multifarious *ceteris paribus* clauses. As Gauker (2003, p. 240) notes, no philosopher or cognitive scientist has succeeded in formulating adequate laws linking mental states to each other and to the observable circumstances that lead to them or to which they lead. But, if providing accurate descriptions of mental facts explains why we interpret each other, then one would think such psychophysical laws would be easier to formulate. This is a version of the computational tractability problem raised above for the theory–theory. Presumably, one of the main reasons we need accurate descriptions of mental states is to coordinate effectively with our fellows. But such coordination typically occurs rapidly and seamlessly, in dynamic, complex environments. If this required first formulating accurate descriptions of all relevant mental states, it would constitute an intractable mystery, given the challenges of computing over the complex *ceteris paribus* laws linking mental states to their observable triggers and consequences.

This argument may seem a little too quick. After all, why think that versions of SP like the theory–theory need be committed to the view that interpretation is implemented via a theory-like representational format, consisting of law-like

[2] See McGeer (2015, p. 266, 271) for a similar point.

generalizations linking mental states to each other and their observable triggers and consequences? And, surely, we *are* aware of some such generalizations, at least as qualified by *ceteris paribus* clauses. For example, other things being equal, humiliation causes resentment, heartbreak causes grief, pride causes disdain, etc. Of course, no such generalizations are exception-less, but they might support heuristics that are correct enough of the time to yield successful description and prediction.[3] It is important to remember here that my target consists in classical forms of SP, like the so-called "theory–theory". These accounts explicitly liken the representational format that underlies interpretation to scientific theories consisting of law-like generalizations linking mental states to each other and their observable triggers and consequences. Furthermore, although it is true that folk psychology is replete with clichés encoding generalizations linking mental states to each other, like humiliation and resentment, for it to support accurate behavioral prediction, generalizations linking mental states to their *observable triggers and consequences* are more important. And there is good reason to doubt that the folk know any *such* generalizations, because there is good reason to doubt that there *are* any such generalizations to be known. For example, Barrett (2012, 2017) provides strong evidence of extreme variability in the behavioral expression of emotions, e.g., via facial expressions. And cross-cultural studies of social cognition provide strong evidence of variability in the behaviors associated with many different kinds of mental states (Lillard 1998; Vinden 1999). The latter evidence suggests that, to the extent that there are reliable generalizations linking mental states to their observable triggers and consequences, they are *products of* culturally mediated regulative pressures. Hence, it is likely that the descriptive and predictive functions of high-level interpretation presuppose its regulative functions. Of course, sub-personal, unreflective, and nonlinguistic forms of interpretation might avoid the holism problem thanks to representational formats that are radically unlike scientific generalizations. This might make possible descriptive and predictive functions that are independent of regulative functions. But such possibilities are not my target here; my objection is to classic forms of SP.[4]

A third reason to doubt that the main function of interpretation is descriptive is that we have a variety of alternative means to predict behavior in coordinative contexts, which appear more computationally frugal than interpretation in terms of mental states. We can generalize from past behavior in similar contexts; we can predict on the basis of teleology, i.e., by identifying the transparent goal of a behavior; we can use our knowledge of norms and social roles to predict

[3] I thank Richard Moore for the objection and the examples.

[4] What about MP? Recall that one of its main advantages over SP is that it does not require knowledge of generalizations linking mental states to their observable triggers and consequences. It is true that MP versions of SC do not succumb to this objection. However, they face another problem: given inter-individual cognitive and motivational variability, MP, *on its own*, cannot explain the reliability of interpretation. This is a problem because, if interpretation is not reliable, it is hard to see how it could perform descriptive and predictive functions. Of course, as many proponents of MP argue, perhaps there is sufficient inter-individual similarity to support reliable interpretation (Goldman 1989). However, the evidence from Barrett (2017) regarding emotions, and from Lillard (1998) and Vinden (1999) regarding cross-cultural variation, suggests that, to the extent that there is such inter-individual similarity, it is the product of regulative functions of interpretation.

behavior, etc. (Andrews 2008; Maibom 2007). Given that there are these other, less computationally costly means of predicting behavior in support of coordinative success, why would we waste time and energy on the epistemically fraught task of constructing accurate descriptions of relevant mental states in coordinative tasks? Of course, it is undeniable that we do routinely engage in mental state attributions. But given their unsuitability to constructing accurate and timely mental state descriptions compared to computationally more frugal alternatives, it is likely that their aims are often different. In particular, there is evidence that mental state attribution is often used to justify rather than to predict behavior (Malle et al. 2007), which would make sense if regulation were an important function of mental state attribution, as justification aims to show that behavior conforms to norms. Indeed, the very property that makes mental state attribution ineffective at descriptive functions, i.e., holism, can explain its efficacy at justification (Zawidzki 2013, Chapter 7). Because any mental state is compatible with any observable behavior or circumstance, given appropriate adjustments to other mental states, it is always possible to use mental state attribution to show that behavior that *apparently* contravenes a norm actually conforms to it. For example, one can justify misinforming a partner about the location of some item by appeal to a blamelessly formed false belief.

A fourth and final reason to doubt that the functions of interpretation are primarily descriptive concerns the role in coordination it is assumed to play. Typically, it is assumed that predicting behavior on the basis of accurate mental state attributions constitutes the main function of interpretation in enabling coordination (Lewis 1969). Two can jointly plan to hunt stag rather than each hunting hare alone because each can predict when and where the other will be, and what the other will do, based on an accurate description of the other's mind. But as is clear from the game theoretic structure of coordinative dilemmas like the "Stag Hunt", accurate mental state attributions appear insufficient to explain successful coordination (Skyrms 2004). The reason is that potential interactants' decisions are contingent on each other. One will choose to hunt stag only if one knows that one's partner will show up to help; otherwise one's time is better spent hunting hare alone. But one's partner is in the same situation. So, each cannot decide to hunt stag until they know that the other has so decided. But when they read each other's minds accurately all they know is that their partner will hunt stag *if they do*, not whether or not they will hunt stag. So accurate mindreading seems to be of no help here. Hence, playing an important role in coordination is *not* a good reason to assume mental state attribution aims primarily at accurate descriptions.

In fact, there is good reason to think that the challenges of largescale coordination on cooperative projects are better met with regulative social practices than with accurate descriptions of mental states. For example, if there are norms governing interactions such that certain communicative acts count as commitments to play roles in future cooperative endeavors, on pain of sanctions, then coordinative dilemmas like the Stag Hunt can be avoided precisely because such norms obviate the need for accurate mental state attributions: cooperation partners can simply assume that each will play their normatively sanctioned role, if communities have effective sanctioning practices.

Such public communication systems and associated norms constitute what Daniel Dennett calls "cognitive tools" (2014): they are culturally transmitted, external structures and practices that simplify a cognitive domain in ways that make it tractable by our limited, internal cognitive resources. However, many such *socio*-cognitive tools have a unique property that distinguishes them from the kinds of cognitive tools on which Dennett focuses, e.g., technologies like calculators, maps, and number systems. Non-social cognitive tools function by enhancing their users' *cognitive agency*. For example, a human being who masters Arabic numerals thereby becomes an enhanced cognitive agent, e.g., she can now engage in mathematical calculations that are otherwise inaccessible to her. *Socio*-cognitive tools, like public communication systems and associated norms, in addition to transforming the cognitive agency of their users, also turn them into new kinds of cognitive *objects*. A person who learns English and certain promise keeping norms has her cognitive *agency* enhanced because she is now able to predict and coordinate with members of communities of English-speakers. But she herself is also, by the same token, transformed as a cognitive *object*: members of communities of English-speakers can now use *their* enhanced cognitive agency on her; she is far more predictable by English-speakers than she was prior to learning English.[5]

It is not surprising that this Janus-faced property, i.e., transforming users as cognitive agents *and* cognitive objects, distinguishes many socio-cognitive tools from non-social cognitive tools. The reason is that, when it comes to social domains, human beings are both cognitive agents and cognitive objects. It is likely that such socio-cognitive tools are key to understanding human evolution, because what most sets us apart from our closest non-human cousins is our capacity for coordination on largescale, cooperative projects, even with individuals of whom we have little personal knowledge (Sterelny 2012). Socio-cognitive tools make this possible by shaping us into the kinds of socio-cognitive objects that are more predictable by the kinds of socio-cognitive agents into which those same tools shape us as well.

If this is on the right track, then the key to our impressive capacities to predict and coordinate with each other is effective regulation, not accurate mental state attribution; yet another reason to suspect regulation is one of the key functions of interpretation. But how, precisely, does interpreting each other's behavior in terms of mental states contribute to the important regulative functions of socio-cognitive tools, like public communication systems and associated norms? On most accounts, such socio-cognitive tools presuppose descriptively accurate mental state attributions (Grice 1975; Lewis 1969; Sperber and Wilson 2002; Scott-Phillips 2015). However, such accounts face the sorts of challenges for the descriptive picture raised above. Fortunately, there are other approaches that envision a different relationship between mental state attributions and socio-cognitive tools like public communication systems and associated norms. On these views, person-level, reflective, linguistically expressible interpretation in terms of high-level metacognitive concepts,

[5] We see a similar dynamic in games and sports. By learning chess, one not only gains new tools for understanding and predicting other chess players, one also becomes easier to understand and predict by other chess players (McGeer 2015, pp. 261–262).

like the propositional attitudes is about situating those with whom one coordinates and communicates in a space of reasons that further enables regulation of behavior (Brandom 1994; Zawidzki 2013; Geurts 2019). Although it is likely that socio-cognitive tools require complex, sub-personal socio-cognitive states to get off the ground (Zawidzki 2013, Chapters 5 and 6; Moore 2017, 2018), person-level interpretive practices, like propositional attitude attribution, are artefacts of, rather than conditions on socio-cognitive tools like public communication systems and associated norms. Explaining how this can be the case requires reconceptualizing the relationship between metacognition and social cognition in ways that move beyond the traditional MP versus SP dichotomy. It is to this that I now turn.

4 Procedural versus conceptual metacognition

The idea I wish to propose and explore is that the person-level, metacognitive concepts we use to interpret our own and others' behavior are themselves socio-cognitive tools that we acquire in ontogeny from our ambient cultures, and that have culturally evolved through human prehistory and history to enable largescale coordination on cooperative projects. By mastering linguistically expressed metacognitive concepts, we transform ourselves as socio-cognitive agents and objects, gaining new tools for interpreting ourselves and others, and simultaneously being made easier to interpret in terms of these tools.

As an analogy, consider concepts of social roles, which clearly function like this. Cultures and languages are repositories of concepts like "parent", "spouse", "teacher", "pupil", "chief", "priest", "citizen", etc., which we are socialized to inhabit. Such concepts, and the socially instituted rules for their use, are clearly socio-cognitive tools that transform us as cognitive agents and objects. By learning what an early 21st-Century North American parent is, I gain a means of interpreting myself and others, in ways that make us more predictable to each other, enhancing our coordinative potential. As for languages more broadly, the functionality of such social role concepts is a frequency-dependent phenomenon: they work only to the extent that they are widespread among those with whom one is likely to interact. For example, I can use the concept "parent" to predict and make myself more predictable to others, only to the extent that they have mastered roughly the same concept, both to regulate their own behavior and to predict mine. The fewer others share this concept with me, the less useful it is to my coordinative projects.

My suggestion here is that person-level, linguistically expressible metacognitive concepts function in exactly the same way. For example, to believe that p is to assume the social role of a believer-that-p; to desire that q is to assume the social role of a desirer-that-q; to grieve that r is to assume the social role of a griever-that-r, etc. By assuming such social roles, one becomes more easily predictable by others wielding the same concepts. As with more typical social roles, the functionality of such metacognitive concepts depends on their prevalence: the more that my fellows use concepts like belief, desire, and grief to interpret their own and others' behavior, including to shape themselves to respect behavioral expectations associated with these concepts, the more useful such metacognitive concepts are at enhancing coordination, by enhancing my predictive capacities, i.e., my capacities qua socio-cognitive agent, and by enhancing my predictability, i.e., my properties qua socio-cognitive object.[6]

Although this is not the typical, causal-functionalist understanding of metacognitive concepts, it is consistent with neo-pragmatist, normative functionalism, e.g., Brandom's (1994) proposal that propositional attitude attributions function as expressions of discursive commitments, and can be analyzed in terms of their deontic relations to other commitments and entitlements. On this view, to say that one believes that p is to express commitment to p, and thereby incur certain entitlements (e.g., to propositions consistent with p) and obligations (e.g., to provide evidence that p). To attribute a belief that p to another is to assert that they are bound by the same entitlements and obligations. To master these concepts is to master the entitlements and obligations that they involve, i.e., to shape oneself into the kind of person that can respect these entitlements and obligations, and to expect others to do the same. Of course, a Brandom-style analysis works better for some metacognitive concepts, e.g., belief, than for others, e.g., emotion. However, there are vibrant empirical paradigms that support social constructivist accounts of emotion that bear some similarities to Brandom's account of discursive commitment. For example, Barrett (2017) argues that full-blown, person-level emotion concepts are culturally-specific glosses on interoceptively detected physiological states like arousal, channeling these initially indeterminate stimuli into culturally familiar behavioral forms. In other words, cultures shape their members to interpret interoceptively detected physiological states in terms of linguistically expressed emotion categories that encode culturally-specific behavioral expectations.[7] So, although expressing, e.g., sadness, does not necessarily involve entitlements and obligations, it does involve culturally specific behavioral dispositions one is socialized to enact. These are, of course, highly controversial accounts of metacognitive concepts, and I do not have the space here to defend them in detail. The point is just that my proposal about person-level, metacognitive concepts is a natural ally to viable, relevant paradigms in philosophy and psychology. To the extent that this proposal provides a superior account of the relation between metacognition and social cognition than either MP or SP, it constitutes one source of evidence in favor of these paradigms.

The first part of my proposal, therefore, is the following. On the regulative conception of metacognition, person-level, linguistically-expressible metacognitive concepts function as socio-cognitive tools, transforming us as socio-cognitive agents and objects, thereby making us more predictable to each other, and enhancing our capacities for coordination. Contrary to MP, we do not use metacognitive concepts to first learn facts about our own minds and then project these onto others in order to better predict them. Contrary to SP, we do not use metacognitive concepts to first

[6] For a congenial discussion, see McGeer (2015, pp. 265–267).

[7] For an anticipation of this idea, see Nietzsche (1881/1997, pp. 26–27).

learn facts about the hidden causes of publicly observable behavior, and then use this knowledge to predict our own and others' behavior. Rather, we use metacognitive concepts to shape our own and others' behavior in ways that make it more predictable in terms of those concepts. But this raises the question of how we are able to do this. The second part of my proposal constitutes the beginnings of an answer: besides our person-level metacognitive *concepts*, we also have sub-personal metacognitive *skills*, or what Proust (2013) calls "procedural metacognition", that enable us to control and regulate our cognitive states in ways that allow us to respect the socially determined rules governing our person-level metacognitive concepts.[8]

Proust introduces the concept of procedural metacognition in the context of making sense of some very robust empirical results concerning non-human metacognition. In a typical paradigm, a nonhuman subject, like a rhesus macaque, is given a visual recall task (Kornell et al. 2007). The subject is first shown a sequence of photographs, followed after a delay by the presentation of an array of photographs, of which only one was shown in the earlier sequence. The subject must then select a photograph from the array. If the subject selects the photograph that was shown before, she receives a food reward; otherwise she receives no reward. Next, a neutral response option is added: the subject is permitted to abstain from selecting a photograph from the array. If the subject abstains, she gets a lesser food reward than she would get if she selected the correct photograph. The reasoning behind this paradigm is the following. If the subject is certain about having seen one of the photographs in the array, then she will select it, since a correct response offers the highest reward. However, if the subject is uncertain, the she will abstain, because a lesser reward is still better than no reward for a false response. But, goes the thinking, certainty and uncertainty are metacognitive judgments: to assess certainty concerning a judgment about a past visual stimulus, one must represent one's own visual memories. Hence, if rhesus macaques, for example, systematically abstain from judgments for stimuli which are harder to remember, this shows that rhesus macaques represent their own visual memories. It is now well established that rhesus macaques and many other nonhuman animals do display this pattern of results in such experimental paradigms.

There is far less consensus about how to interpret these experimental results, however. Most of the ingenious experimental design has gone into distinguishing between behaviorist and cognitivist interpretations of the results (Ibid). Experimenters have shown conclusively that the discriminations made in such experiments cannot be accounted for solely in terms of the relative rewards of perceptually available distinctions among the test stimuli: stimuli eliciting similar responses have nothing perceptual in common, and so must be represented abstractly by the nonhuman subjects. However, as Carruthers argues (2009), this is not sufficient to show that nonhumans can represent their own cognitive states, that they are capable of metacognition. Perhaps they are simply representing all the relevant stimuli, triggering a competition for behavioral response.

[8] See McGeer (2015, pp. 261–267) for an insightful discussion of such "folk psychological know-how or expertise".

Representations of stimuli that are similar enough to those seen before have enhanced activation, and so are more likely to win this competition for behavioral response, and be selected. When two stimuli are both similar enough to one of the past stimuli, this yields hesitation, *not* because the subject is metarepresenting her own representations of these stimuli and comparing them to her visual memories, but simply because their activations are equally enhanced, and so the competition for behavioral response is evenly matched. When the option to abstain is introduced, the subject learns to select it in response to such hesitation. The results can be explained without assuming the nonhuman subjects can represent their own representations, or that they are capable of metacognitive judgments.

This debate is partly driven by differences in background commitments involving SP and MP. For example, Carruthers is skeptical that nonhuman animals are capable of metacognition because he endorses SP, and hence thinks such self-directed metacognition is derivative of social cognition (Carruthers 2011). Since there is little evidence that nonhuman animals use metacognition to understand the social domain, it follows that neither are they capable of self-directed metacognition. Many of those who defend the metacognitive interpretation of these experimental results, on the other hand, explicitly endorse MP: they argue that meta-cognition was first used, in phylogeny, for non-social purposes, and then adapted to social ones (Couchman et al. 2009). Proust (2013) defends a third alternative which forms the basis of my proposal in this paper. She interprets the non-human metacognition studies as evidence for a non-conceptual, procedural metacognitive capacity: a subject's ability to control its own mental states without having to represent them, i.e., without having metacognitive *concepts*. It should immediately be clear why Proust's idea is congenial to my proposal here. If it is possible to shape one's own cognitive states without using metacognitive concepts, then such concepts can be understood as *products of* rather than as *necessary for* such metacognitive skills. This would clear the way for conceiving of metacognitive concepts as socio-cognitive tools, as proposed above.

Proust (2013) proposes two persuasive arguments for interpreting nonhuman metacognitive competence in terms of non-conceptual, procedural metacognition; however, the proposal also faces a serious challenge. Proust's first argument draws on the same premise as Carruthers' skepticism about nonhuman metacognition: there is little evidence that nonhumans use metacognition to navigate the social domain. However, unlike Carruthers, Proust concludes that this shows only that nonhumans do not have metacognitive *concepts*; it does not show that they are incapable of procedural metacognition. For Proust (2013, p. 30), nonhuman limitations on using metacognitive concepts to navigate the social domain count against the hypothesis that nonhumans can use such concepts to control their own cognitive states because any conceptual competence must respect Evans's "generality constraint" (1982): roughly, a subject has a concept P if she can entertain thoughts consisting of the combination of P with any other concept she has. But, because there is little evidence that nonhumans can conceive of their conspecifics as subjects of mental states, this means that they do not represent their own mental states conceptually, since they cannot entertain thoughts consisting of the combination of a metacognitive representation with a conspecific representation. For example, they cannot think of their siblings that they are sad, and hence, do not have the concept of sadness.

Proust's second argument for interpreting nonhuman metacognitive competence in terms of non-conceptual, procedural meta-cognition appeals to the fine-grained, seamless, dynamic control provided by this competence. If nonhuman subjects had to first meta-represent, in a separate, metacognitive faculty, the representations involved in some first-order cognitive process, before gaining appropriate control over that process, it would be a mystery how they could gain such seamless, dynamic control over the process. For example, in the experimental paradigm discussed above, in order to master the option of abstaining from selecting a photograph, rhesus macaques would first have to meta-represent their visual memories, and apply some sort of reliability metric to them, in order to gauge their uncertainty. Proust argues that it is far more plausible that they are simply driven by "noetic feelings", e.g., uncertainty, that arise as by-products of first-order cognitive processes, like the ease with which a visual memory is retrieved (Proust 2013, 14); there is no need to meta-represent the visual memory in addition. Although such processes are not meta-representational, Proust argues that they count as metacognitive, since they involve the intelligent manipulation and control of first-order cognitive states. Hence, they constitute a kind of non-conceptual, procedural metacognition.

Langland-Hassan (2014) raises an important challenge to Proust's proposal, however: it risks an over-proliferation of metacognition. If metacognitive concepts do not constitute the distinctive mark of a metacognitive process, then what does? As Langland-Hassan points out, without this way of distinguishing metacognitive processes, there is a risk that

> metacognition occurs wherever a cognitive process has the function of controlling, monitoring, or calibrating another cognitive process. Many agree that the most basic mechanisms governing action and perception involve subconscious prediction and comparison processes that fulfill these criteria… Suddenly fly-swatting and jump-roping become metacognitive events. (ibid, 723)

But I think this challenge is overstated. We can conceive of procedural metacognition as admitting of degrees, depending on how much control over first-order cognitive processes it provides. Perhaps there are relatively trivial varieties of procedural metacognition involved even in fly-swatting. This does not rule out the possibility of more sophisticated varieties that still do not require metacognitive concepts. What matters is the degree of control one cognitive process provides over another.

I am now in a position to make more explicit the role I think procedural metacognition plays in helping humans master linguistically expressible, person-level, metacognitive concepts that function as socio-cognitive tools. In the same way that non-human and human subjects can learn to flexibly control responses to visual memories via the noetic feeling of uncertainty, implemented in non-conceptual, procedural, metacognitive capacities, human subjects can learn to flexibly control *socially relevant* first-order cognitive states via other kinds of feelings. For example, feelings like shame, guilt, and embarrassment might play roles in non-conceptual, procedural metacognitive processes that control first-order, socially relevant cognitive processes, such as those triggering behavioral responses to social situations. Even simple failures to successfully predict others' behavioral responses could drive such metacognitive adjustments without requiring metacognitive concepts, i.e., the capacity to represent the states being adjusted *as cognitive*. There is some evidence, for example, that relevant areas of the human brain treat simple failures to conform to group judgments, e.g., regarding facial attractiveness, as error signals that lead to modifications of the neural mechanisms giving rise to these judgments in a conformist direction (Klucharev et al. 2009). Here, there is clearly a kind of metacognitive control driven by social factors, but no reason to posit metacognitive concepts.

Assuming that human brains are distinguished from those of other animals by the richness and flexibility of such socially-modulated, non-conceptual, metacognitive processes, we can explain how we come to master the use of metacognitive concepts as socio-cognitive tools. Our dispositions to use words that express such concepts are shaped via reactions by those to whom we express the concepts. For example, we learn how we are supposed to use the word "sad", in labelling our own and others' behavior, by gauging the feedback that others, often unwittingly, provide: failures to predict feedback, or negative feedback eliciting shame or embarrassment, guide subpersonal, non-conceptual and procedural, metacognitive adjustments, tuning our dispositions to match those of the ambient culture. Similar processes can explain how we learn behavior appropriate to situations that we or others have labelled as "sad", enabling the self-regulating use of this metacognitive concept and socio-cognitive tool. We thereby both become easier to interpret by those with whom we interact, and learn how to better interpret them.[9] I now turn to a brief assessment of how this view of the relationship between metacognition and social cognition compares to MP and SP, relative to the explanatory virtues and vices discussed above.

5 Conclusion: the virtues of the regulative alternative to MP and SP

As I explained in Sect. 2, MP and SP display a pattern of complementary explanatory virtues and vices. MP seems more intuitive or phenomenologically plausible than SP, but SP avoids positing a mysterious notion of introspective access. MP can better explain the evident epistemic asymmetries between self- and other-directed metacognition, but SP can better explain the fact that we use the same concepts in self- and other interpretation. MP has a better account of how other interpretation

[9] This is not to suggest that such socially inflected, procedural metacognition is *sufficient* for learning how to use words like "sad". All language-learning requires, in addition, enormously complex socio-cognitive machinery that enables the kinds of pragmatic inferences by means of which language learners infer the intended messages of their interlocutors. Richard Moore, whom I thank for raising this point in response to an earlier draft, has argued (2017, 2018) that such Gricean mechanisms needn't involve sophisticated metacognitive concepts, of the kind we use in person-level, reflective, and language-involving interpretation. For example, he argues that non-humans and human infants can engage in Gricean forms of communication without a concept of belief or the ability to attribute recursively nested mental states.

can be quick, seamless, and computationally tractable, but SP has a better account of how other interpretation can be reliable. SP makes better sense of some empirical results, like the ontogeny of mental state concepts, while MP makes better sense of other empirical results, like prevalent egocentric biases, such as the "curse of knowledge". How does the alternative, regulative conceptualization of the relation between metacognition and social cognition fare relative to these explanatory virtues and vices?

Here is the alternative in a nutshell. Person-level, linguistically expressible, metacognitive concepts are socio-cognitive tools that individuals acquire from their cultures, and that transform them as cognitive agents and cognitive objects, making them better at predicting others and easier to predict by others. Thus, they have both regulative and descriptive/predictive functions, but the latter depend on the former: we can use our metacognitive concepts to describe and predict each other's behavior because we have used them to regulate our own and each other's behavior in ways that make it easily describable and predictable in their terms. At a sub-personal level, mastering these socio-cognitive tools is made possible via non-conceptual, procedural metacognition, which consists in adjusting socially relevant behavioral dispositions in response to social feedback, in the form of prediction errors, or behavior triggering emotions like embarrassment, shame, and guilt.

This conception of the relation between metacognition and social cognition combines the explanatory virtues and avoids the explanatory vices of MP and SP. Most obviously, it can explain both the epistemic asymmetries between self- and other-directed metacognition, and the fact that we use the same concepts for ourselves and others. Key to this explanation is the idea that our metacognitive concepts are socio-cognitive tools. This means that they depend for their reliability as descriptions on their widespread use in self-regulation: it is because we use the same concepts to regulate ourselves that these concepts are so reliable at describing ourselves. The most typical regulative uses of our metacognitive concepts are self-directed. As McGeer puts it:

> I cannot make it the case that you behave in ways coherent with what I say you hope, desire, or fear any more than I can make it the case that the world is a certain way by announcing how (I think) it is; but I can and do govern my own actions in ways that fit with the claims I make about myself. If so-called "knowledge" of our own minds thus consists largely of claims we have both made and acted in light of, it is no surprise that such "knowledge" is peculiarly authoritative. (1996, p. 507)

This explains the epistemic asymmetries. But these metacognitive concepts can support coordination only if they can also be used descriptively, to anticipate the behavior of others. It is only because we all use roughly the same concepts to self-regulate and predict each other that they can play the roles of socio-cognitive tools that enhance coordination.

The trade-off between reliability and computational tractability is also neatly avoided on the regulative alternative to MP and SP. Conceiving of metacognitive concepts as socio-cognitive tools shows how metacognition can be both computationally tractable and reliable. Because the same metacognitive concepts are used both in self-regulation and in prediction by groups of likely interactants, we know that the concepts we use to interpret and regulate ourselves will also serve to predict those with whom we interact. The kinds of socially-driven, non-conceptual, procedural, meta-cognitive capacities described above shape populations of likely interactants to self-regulate and predict others using mutually-calibrated metacognitive concepts, thereby ensuring the reliability and computational tractability of these socio-cognitive tools.

The regulative alternative to MP and SP can also avoid some hard choices when it comes to explaining empirical findings. With SP, it predicts that children will learn to apply metacognitive concepts to themselves and others at roughly the same time. The reason is that mastering a socio-cognitive tool requires mastering both self-directed regulative uses and other-directed predictive uses. On the regulative alternative, metacognitive concepts encode norms of behavior, and one does not understand a norm unless one knows how it applies both to oneself and others, in both regulative and descriptive uses. At the same time, the regulative alternative can explain the prevalence of egocentric biases, like the "curse of knowledge". If one takes one's self-interpretations as norms for rather than descriptions of behavior, then one will of course assume that others live up to the same norms.

Finally, the regulative alternative to MP and SP goes some way to explaining our intuitions regarding metacognition without positing a mysterious introspective mechanism. It seems intuitive that we first know our own minds and then project this knowledge onto others *not* because we have mysteriously transparent access to our own cognitive states but, rather, because descriptive uses of metacognitive concepts derive from their regulative uses, and the latter are typically self-directed, while the former are typically other-directed. Other-interpretation seems intuitively to derive from self-interpretation because description in terms of mental states relies on regulation in terms of mental states, and, as McGeer (1996) argues, the latter is much more characteristic of self- than other-interpretation.

However, it is worth pointing out that this is not a complete vindication of commonsense phenomenology or intuition. There is no denying that interpretation, whether of self or other, does not *feel* like a regulative act. This is the one sense in which both MP and SP are explanatorily superior to the regulative alternative. They begin with what Hutto calls the "spectatorial" assumption that interpretation aims, in the first instance, to describe independent facts about the mental states of self and others. This is clearly the commonsense view. In fact, it is arguable that interpretation succeeds so well at regulation precisely because it conceals itself as description. Similar claims have been made about socio-politically fraught categories, like race and gender (Haslanger 2012). I do not, however, regard this as problematic for the regulative alternative to MP and SP. There is no view that can simultaneously satisfy all commonsense intuitions, as well as scientific constraints like consistency with empirical results and the broader, scientific world-view. Contradicting the commonsense assumption that interpretation aims primarily at description, not regulation, is arguably not even an explanatory vice: all it shows is that the regulative conception of the relation between metacognition and social cognition has much in common with other influential re-conceptualizations of our interpretive practices, e.g., those of Nietzsche, Marx and Foucault, according to which they generate ideologies masquerading as descriptions of natural facts.

In conclusion, the metacognition-as-primary and the social-cognition-as-primary views share the spectatorial conception of our interpretive practices: according to both views, mental state attribution aims primarily to describe an independently constituted domain of mental facts. I have argued that if we reject the spectatorial conception in favor of the regulative conception of our interpretive practices—a theoretical move for which there are persuasive, independent reasons, a new conception of the relationship between metacognition and social cognition becomes available. This new conception is centered on the idea that person-level, linguistically expressible, metacognitive concepts are socio-cognitive tools that enhance human coordination by shaping us into better cognitive agents and cognitive objects. We master these socio-cognitive tools thanks to a non-conceptual, procedural, metacognitive capacity for controlling and regulating our own cognitive states in response to social feedback. This reconceptualization of the relation between metacognition and social cognition combines the complementary explanatory virtues and avoids the complementary explanatory vices of the metacognition-as-primary and the social-cognition-as-primary views.

Acknowledgements I thank Richard Moore for extremely helpful, extensive comments on an earlier draft.

References

Andrews, K. (2008). It's in your nature: A pluralistic folk psychology. *Synthese, 165,* 13–29.

Barrett, L. (2012). Emotions are real. *Emotion, 12,* 413–429.

Barrett, L. (2017). *How emotions are made*. New York: Houghton, Mifflin, Harcourt.

Brandom, R. (1994). *Making it explicit*. Cambridge, MA: Harvard University Press.

Butterfill, S., & Apperly, I. (2013). How to construct a minimal theory of mind. *Mind and Language, 28,* 606–637. https://doi.org/10.1111/mila.12036.

Carruthers, P. (1996). Simulation and self-knowledge: A defence of the theory-theory. In P. Carruthers & P. Smith (Eds.), *Theories of theories of mind*. Cambridge, UK: Cambridge University Press.

Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences, 32,* 121–182.

Carruthers, P. (2011). *The opacity of mind*. Oxford: Oxford University Press.

Couchman, J., et al. (2009). Metacognition is prior. *Behavioral and Brain Sciences, 32,* 142.

Davidson, D. (1984). *Inquiries into truth and interpretation*. Oxford: Oxford University Press.

Dennett, D. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Dennett, D. (2014). *Intuition pumps and other tools for thinking*. New York: W. W. Norton & Company.

Descartes, R. (1641/1993). *Meditations on first philosophy.* Indianapolis: Hackett.

Evans, G. (1982). *The varieties of reference*. Oxford: Oxford University Press.

Gauker, C. (2003). *Words without meaning*. Cambridge, MA: MIT Press.

Geurts, B. (2019). Communication as commitment sharing: Speech acts, implicatures, common ground. *Theoretical Linguistics, 45,* 1–30.

Goldman, A. (1989). Interpretation psychologized. *Mind and Language, 4,* 161–185.

Goldman, A. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. New York: Oxford University Press.

Golombok & Fivush. (1994). *Gender development*. Cambridge, UK: Cambridge University Press.

Gopnik, A., & Astington, J. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Development, 59,* 26–37.

Gopnik, A., & Meltzoff, A. (1996). *Words, thoughts, and theories*. Cambridge, MA: MIT Press.

Gordon, R. (1986). Folk psychology as simulation. *Mind and Language, 1,* 158–171.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (41–58).

Hacking, I. (1995). The looping effects of human kinds. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Symposia of the Fyssen foundation. Causal cognition: A multidisciplinary debate* (pp. 351–394). New York: Oxford University Press.

Hacking, I. (1998). *Mad travelers*. Charlottesville: University of Virginia Press.

Haslanger, S. (2012). *Resisting reality: Social construction and social critique*. Oxford: Oxford University Press.

Hume, D. (1739/1978). *A treatise of human nature*. P. H. Nidditch (Ed.), 2nd Ed. Oxford: Oxford University Press.

Hutto, D. (2008). *Folk psychological narratives*. Cambridge, MA: MIT Press.

Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., & Fernández, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron, 61,* 140–151.

Kornell, N., Son, L., & Terrace, H. (2007). Transfer of metacognitive skills and hint seeking in monkeys. *Psychological Science, 18,* 64–71.

Langland-Hassan, P. (2014). Unwitting self-awareness? *Philosophy and Phenomenological Research, 89,* 719–726.

Lewis, D. (1969). *Convention*. Oxford: Blackwell.

Lillard, A. (1998). Ethnopsychologies: Cultural variations in theories of mind. *Psychological Bulletin, 123,* 3–32. https://doi.org/10.1037/0033-2909.123.1.3.

Maibom, H. (2007). Social systems. *Philosophical Psychology, 20,* 557–578.

Malle, B., Knobe, J., & Nelson, S. (2007). Actor-observer asymmetries in explanations of behavior: New answers to an old question. *Journal of Personality and Social Psychology, 93,* 491–514.

Mameli, M. (2001). Mindreading, mindshaping, and evolution. *Biology and Philosophy, 16,* 597–628.

McGeer, V. (1996). Is 'self-knowledge' an empirical problem? Renegotiating the space of philosophical explanation. *Journal of Philosophy, 93,* 483–515.

McGeer, V. (2001). Psycho-practice, psycho-theory and the contrastive case of autism. *Journal of Consciousness Studies, 8,* 109–132.

McGeer, V. (2007). The regulative dimension of folk psychology. In D. Hutto & M. Ratcliffe (Eds.), *Folk psychology reassessed*. Dordrecht: Springer.

McGeer, V. (2015). Mind-making practices: The social infrastructure of self-knowing agency and responsibility. *Philosophical Explorations, 18,* 259–281. https://doi.org/10.1080/13869795.2015.1032331.

Moore, R. (2017). Gricean communication and cognitive development. *The Philosophical Quarterly, 67,* 303–326. https://doi.org/10.1093/pq/pqw049.

Moore, R. (2018). Gricean communication, language development, and animal minds. *Philosophy Compass*. https://doi.org/10.1111/phc3.12550.

Morton, A. (2003). *The importance of being understood*. London: Routledge.

Nichols, S., & Stich, S. (2003). *Mindreading*. Oxford: Oxford University Press.

Nietzsche, F. (1881/1997). *Daybreak: Thoughts on the prejudices of morality*(2nd ed.). Cambridge, UK: Cambridge Texts in the History of Philosophy.

Nix, S., Perez-Felkner, L., & Thomas, K. (2015). Perceived mathematical ability under challenge: A longitudinal perspective on sex segregation among STEM degree fields. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2015.00530.

Proust, J. (2013). *The philosophy of metacognition*. Oxford: Oxford University Press.

Russell, B. (1948/2009). *Human knowledge: Its scope and limits*. London: Routledge Classics.

Scott-Phillips, T. (2015). *Speaking our minds: Why human communication is different, and how language evolved to make it special*. London: Palgrave.

Searle, J. (1979). *Expression and meaning*. Cambridge: Cambridge University Press.

Sellars, W. (1956). Empiricism and the philosophy of mind. *Minnesota Studies in the Philosophy of Science, 1,* 253–329.

Skyrms, B. (2004). *The stag hunt and the evolution of social structure*. Cambridge, UK: Cambridge University Press.

Sperber, D., & Wilson, D. (2002). Pragmatics, modularity, and mind-reading. *Mind and Language, 17,* 3–23.

Sterelny, K. (2012). *The evolved apprentice*. Cambridge, MA: MIT Press.

Vinden, P. (1999). Children's understanding of mind and emotion: A multi-culture study. *Cognition and Emotion, 13,* 19–48. https://doi.org/10.1080/026999399379357.

Weiskopf, D. (2005). Mental mirroring as the origin of attributions. *Mind and Language, 20,* 495–520.

Wellman, H. (1990). *The child's theory of mind*. Cambridge, MA: MIT Press.

Zawidzki, T. (2008). The function of folk psychology: Mind reading or mind shaping? *Philosophical Explorations, 11*(3), 193–210.

Zawidzki, T. (2013). *Mindshaping*. Cambridge, MA: MIT Press.