Thinking about past minds: Cognitive science as philosophy of historiography

Forthcoming in *Journal of the Philosophy of History* | Adam Michael Bricker (Turku)

Abstract

This paper outlines the case for a future research program that uses the tools of experimental cognitive science to investigate questions that traditionally fall under the remit of the philosophy of historiography. The central idea is this—the epistemic profile of historians' representations of the past is largely an empirical matter, determined in no small part by the cognitive processes that produce these representations. However, as the philosophy of historiography is not presently equipped to investigate such cognitive questions, legitimate concerns about evidential quality go largely overlooked. The case of mental state representation provides an excellent illustration of this. Representations of past mental states—the thoughts and fears and knowledge and desires of past agents—play much the same evidential role in historiography as in everyday life, serving in the causal explanation of agents' behaviors and supporting normative evaluation of those behaviors. However, we have good reason to suspect that the theory of mind processes that support these representations may be more susceptible to error when deployed in the context of historiography than under everyday conditions. This raises worries about the quality of evidence that theory of mind can provide historiography, worries which require experimental cognitive science to properly address.

Keywords: theory of mind, mentalizing, mindreading, epistemic egocentrism, episodic simulation, epistemology, Collingwood

1. Introduction

The philosophy of historiography is naturally understood as a largely epistemological enterprise.[1] Historiography produces representations of the past, and the philosopher of historiography asks the sorts of questions that an epistemologist could ask about any representations. How do these representations relate to the evidence? In virtue of what are they justified? Do they constitute knowledge? And so on.

Due to this broad structural similarity, the philosophy of historiography is ripe for integration with contemporary mainstream epistemology. Particularly given the real-world significance of its subject matter, the philosophy of historiography would be perfectly at home within the

[1] See, e.g., Aviezer Tucker, *Our knowledge of the past: a philosophy of historiography* (Cambridge, UK: Cambridge University Press, 2004), 2; Jouni-Matti Kuukkanen, "Historiographical Knowledge as Claiming Correctly" in Jouni-Matti Kuukkanen, ed., *Philosophy of History: Twenty-First-Century Perspectives* (London: Bloomsbury, 2021).

world of applied epistemology, which has recently seen high-profile work on topics like evidence in law[2] and political discourse.[3] Such cross-disciplinary interaction could enrich the theoretical tools available to the philosophy of historiography and promote broader awareness of the discipline. This paper, however, isn't really about that. Instead, I want to highlight a much less obvious but equally significant way in which the philosophy of historiography stands to benefit from the theoretical and even methodological input of another field. In slogan form, this paper advocates for a research program that *uses cognitive science to do philosophy of historiography*.

The central observation of my proposal is this—when historiography produces representations of the past, it does so in no small part thanks to the cognitive processes of historians. The nature of these representations generally, and their epistemic profile in particular, will be partially determined by the empirical features of the cognitive processes that produce them. In contemporary epistemology, it is broadly recognized that belief-forming processes play a major role in determining the epistemic profile of the beliefs produced by those processes,[4] and increasing attention is paid to how empirical features of the cognitive processes underlying belief formation impart a particular epistemic character.[5] Within the philosophy of historiography, however, empirical features of human cognition attract much less attention. While it's not unusual to find discussion of processes like inference,[6] this tends to focus on theoretical features of structure, not empirical details of cognitive implementation. Here I'll suggest that the cognitive science of our reasoning about history deserves more attention than it has previously received, sketching one possible future for philosophy of historiography—an integrated cognitive science of historiographical representation.

In advocating for this use of cognitive science as philosophy of historiography, I'll focus specifically on one particular component of human cognition—our capacity to represent the mental states of others. Often referred to collectively as "theory of mind", mental state representation offers an excellent proof of concept in how the findings and methods of cognitive science might be harnessed in the service of philosophy of historiography. On the one hand, extant empirical findings raise significant concerns about the quality of evidence that theory of mind provides when reasoning about the past. While generally reliable in ordinary, everyday contexts, it is likely that the conditions inherent to reasoning about the past exacerbate underlying limitations in our theory of mind capacities.[7] As historians' representations of past figures' mental states play a significant evidentiary role in historiography, the epistemic profile of these mental state representations falls neatly within the purview of the philosophy of historiography. However, the philosophy of historiography itself is not properly equipped to answer these questions, which are thoroughly empirical in nature and require empirical methods to address. Enter the cognitive science of mental state representation. By experimentally investigating the cognitive processing underlying how we

[2] See, e.g., Martin Smith, "When Does Evidence Suffice for Conviction?", *Mind*, 127 (2018): 1193–1218.

[3] See Michael Hannon and Jeroen de Ridder, *The Routledge handbook of political epistemology* (London: Routledge, 2021).

[4] See, e.g., Alvin I. Goldman, "Discrimination and Perceptual Knowledge," *The Journal of Philosophy*, 73 (1976): 771–791; Duncan Pritchard, *Epistemic Luck* (Oxford: Oxford University Press, 2005).

[5] See, e.g., Jonathan Weisberg, "Belief in psyontology," *Philosophers' Imprint*, 20 (2020): 1–27; Adam M. Bricker, "Close Error, Visual Perception, and Neural Phase: A Critique of the Modal Approach to Knowledge," *Theoria*, 87 (2021): 1123–1152.

[6] See, e.g., Kuukkanen, "Historiographical Knowledge as Claiming Correctly".

[7] Moreover, it's likely that similar conditions hold for other disciplines in the humanities. This will come up again at the end of the paper, but otherwise we'll focus specifically on the case of history.

think about past minds, we can gain a deeper understanding of the possible limitations of using mental state representations as evidence in historiography.

To be clear, in this paper I unfortunately won't be doing this kind of experimental investigation. That would require a dedicated research effort, bringing together specialists in both philosophy of historiography and cognitive science. Instead, here I'll set my sights on a much more modest goal—arguing that there is a legitimate need for such a project. There are important questions that fall under the remit of the philosophy of historiography, like the quality of historiographical evidence provided by theory of mind, which the philosophy of historiography isn't presently equipped to answer. In doing so, the paper will go like this: I'll first say a bit on the basics of our human theory of mind capacities (Section 2). What exactly do they do? And how might they go wrong? I'll then sketch some of the evidential roles that mental state representations play in historiography (Section 3). While far from a comprehensive or systematic account, it should be sufficient to highlight the historiographical importance of theory of mind, along with the epistemic worries that this raises. After this, I'll discuss how certain philosophical accounts may be particularly susceptible to the limitations of thinking about past minds, taking Collingwood as an example (Section 4). I'll close with a word on the pragmatic limitations facing the project (Section 5).

2. Thinking about other minds

The ability to represent and track the mental states[8] of others is central to human social cognition. A capacity referred to collectively as "theory of mind",[9] "mentalizing", or "mindreading", we intuitively form representations of what others see and hear, think and believe, want and desire, hope, fear, know, understand, and so on.[10] Depending on the specific mental state, this capacity can be supported by a wide variety of underlying processes which operate on a diverse set of cues and inputs. The representation, for example, of others' simple visual states—what they see in their immediate environments—relies heavily on gaze-tracking.[11] Unsurprisingly, a central component of accurately representing what others see is accurately tracking their line of sight. In contrast, the representations of others' degrees of confidence in their assertions relies largely on prosodic cues like pitch dynamics—roughly, how their assertions sound.[12] For example, broadly speaking, more confident assertions are associated with decreasing pitch whereas less confident assertions are

[8] Note that here I'll understand relations to mental representations which persist through time—such as beliefs and knowledge—to number among paradigm instances of mental states. For more on this representational approach to the mind, see David Pitt. "Mental Representation", in Edward N. Zalta and Uri Nodelman eds., *The Stanford Encyclopedia of Philosophy* (Fall 2022 Edition), especially Section 1.

[9] Note too that the "theory" in "theory of mind" need not be taken literally. While some do view our capacity to represent others' mental states as driven by a quite literal folk theory of the mind—see, e.g., Henry M. Wellman, *Making minds: how theory of mind develops* (Oxford: Oxford University Press, 2014)—this "theory-theory" certainly isn't the only interpretation. See, e.g., Caitlin E. V. Mahy, Louis J. Moses and Jennifer H. Pfeifer, "How and where: Theory-of-mind in the brain," *Developmental Cognitive Neuroscience*, 9 (2014): 68–81. Here I'll try to remain neutral regarding how best to interpret theory of mind capacities.

[10] For introductions, see, e.g., Ian Apperly, *Mindreaders: The Cognitive Basis of "Theory of Mind"* (New York: Psychology Press, 2011); Mahy et al., "How and where: Theory-of-mind in the brain".

[11] See Pascale Michelon and Jeffrey M. Zacks, "Two kinds of visual perspective taking," *Perception & Psychophysics*, 68 (2006): 327–337; Evan Westra and Jennifer Nagel, "Mindreading in conversation," *Cognition,* 210 (2021): 104618, Section 4.

[12] For an overview, see Adam M. Bricker, "I Hear You Feel Confident," *The Philosophical Quarterly*, 73 (2022): 24–43.

associated with increasing pitch.[13] Note that while it may not be the case for the representation of all mental states, the foundations of theory of mind have long developmental and evolutionary histories. The capacity to represent *knowledge* in particular displays especially deep roots, observed in not only adult humans but also human infants and even non-human primates.[14]

Representations of others' mental states play a variety of important roles in everyday life. Perhaps most fundamentally, we use mental state representations in the explanation and prediction of the actions and behaviors of others.[15] The central assumption here is that an agent's mental states will play a primary causal role in her actions. Different mental states cause different kinds of behavior. So by accurately representing her mental states, we can reliably explain what that agent does and predict what she will do. If we observe someone let out a shriek and engage in avoidance behavior upon seeing a spider, we can explain that behavior with the mental state of fear. Or say we watch as a driver blows through a stop sign without as much as slowing down or even turning to look for oncoming traffic. He almost certainly *believed* that he had the right of way and did not *see* that there was a stop sign. Action prediction functions much in the same way. Say I'm meeting a famously punctual colleague for lunch at some restaurant familiar to both of us, and I accordingly represent her as knowing where the restaurant is. On this basis, I predict that she will be able to reach the restaurant without any trouble. She doesn't need any directions or help finding the place and will make it on time. However, if instead the restaurant recently moved, it may be that she believes, incorrectly, that the restaurant is still at the old location. In this case, without being given this new information, she'll likely go to that old location first and thus be late for lunch. I may not consciously realize that I'm thinking about her mental states when I predict whether, without intervention, she'll go to the wrong location. But these representations play a central role in my decision whether to send her directions or an address.

Another important function of mental state representation is that it provides a basis for subsequent normative evaluations of agents' actions and behaviors, particularly what propositions they should *assert* and what propositions they should *act based on*. While it's the subject of considerable debate as to what exact mental state representations these are,[16] it is widely thought that they are either *knowledge states* or *belief states*. On the knowledge norm of assertion, agent S should assert that p only if S knows that p.[17] I should, for example, assert to a group of vegans that a dish I've prepared is vegan only if I know that it is. In contrast, on a belief norm of assertion, S should assert that p only if S believes that p.[18] While there may be some additional epistemic requirements on the belief, such as reasonableness, S need not have knowledge that p in order to assert that p. On such an account, even if the dish

[13] Xiaoming Jiang and Marc D. Pell, "The sound of confidence and doubt," *Speech Communication*, 88 (2017): 106–126.

[14] Jonathan Phillips, Wesley Buckwalter, Fiery Cushman, Ori Friedman, Alia Martin, John Turri, Laurie Santos and Joshua Knobe, "Knowledge before belief," *The Behavioral and Brain Sciences*, 44 (2021): e140.

[15] See, e.g., Chris L. Baker, Rebecca Saxe and Joshua B. Tenenbaum, "Action understanding as inverse planning," *Cognition*, 113 (2009): 329–349; Victoria Southgate and Angelina Vernetti, "Belief-based action prediction in preverbal infants," *Cognition*, 130 (2014): 1–10; John Turri, "Knowledge Attributions and Behavioral Predictions," *Cognitive Science*, 41 (2017): 2253–2261; Evan Westra, "Stereotypes, theory of mind, and the action–prediction hierarchy," *Synthese*, 196 (2019): 2821–2846.

[16] See Matthew Weiner, "Norms of Assertion," *Philosophy Compass*, 2 (2007): 187–195; Jessica Brown, "Knowledge and Practical Reason," *Philosophy Compass*, 3 (2008): 1135–1152.

[17] See, e.g., Timothy Williamson, *Knowledge and its limits* (Oxford: Oxford University Press, 2000), Chapter 11; John Turri, "Evidence of factive norms of belief and decision," *Synthese*, 192 (2015): 4009–4030; John Turri, "Revisiting norms of assertion," *Cognition, 177* (2018): 8–11.

[18] See, e.g., Jennifer Lackey, "Norms of Assertion," *Noûs*, 41 (2007): 594–626.

I've prepared isn't actually vegan, I can permissibly assert that it is, provided that I believe (on a reasonable basis) that it is. Moving on, the knowledge–belief divide for norms of practical reasoning breaks down in much the same way. On the knowledge norm of practical reasoning, S should act based on p only if S knows that p.[19] The CEO of a medical diagnostics company, for example, should act as if her flagship device works—securing contracts to put the device into nationwide commercial use, running diagnostic tests on the device with real patients, and so forth—only if she knows that it actually works. In contrast, a belief norm of practical reasoning might say that S should act based on p only if S (reasonably) believes that p.[20] On such an account, our CEO may permissibly act as if her device works, even if it doesn't, provided that she believes on good reason that it does. Note that for a sizable proportion of ordinary cases, belief and knowledge norms will provide the same results, as they only come apart in cases of epistemically justified false belief (or, perhaps, epistemically justified true belief that still falls short of knowledge). For this reason, as well as a good deal of positive empirical evidence, particularly for folk evaluations[21], here I'll assume knowledge norms for assertion and practical reasoning. Nevertheless, in either case, we observe that our representations of others' mental states play a key role in our normative evaluation of their behavior. Whether or not S's assertions or actions are normatively permissible is not just the function of the content of those assertions and actions but S's mental states in performing them.

Finally, theory of mind also plays an important role in facilitating learning about the world from others. A function largely unique to *knowledge* representations, we can flag others as good informants—reliable sources of information about some topic—by representing them as knowing about that topic.[22] For example, let's say that I want to learn whether Spanish is the first language of everyone in Spain. To gain this information, I can ask someone who I represent as knowing whether Spanish is the first language of everyone in Spain, perhaps an expert on European languages or, more likely, a Spanish person. Upon doing so, I quickly learn that there are a variety of linguistic communities in Spain, and, while Spanish is the first language of most Spanish people, it certainly isn't the only one. Observe that in order to do this, I represent my informant as knowing something that I myself don't know. Referred to as representation from a position of "egocentric ignorance",[23] this is a hallmark of knowledge attribution, facilitating its trademark role in flagging informants. Moreover, observe that knowledge can only play this role because it is factive[24]—knowledge requires truth. When we represent someone as knowing whether something is the case, we represent them as being correct about that something. Contrast this with non-factive states like belief or confidence, which can be held to either truths or falsehoods. I cannot flag someone as a good informant about a topic just because they have beliefs about the topic, no matter how confident they

[19] For example, John Hawthorne and Jason Stanley, "Knowledge and Action," *The Journal of Philosophy*, 105 (2008): 571–590; Turri, "Evidence of factive norms of belief and decision".

[20] For example, Brown, "Knowledge and Practical Reason".

[21] Turri, "Evidence of factive norms of belief and decision"; John Turri, Ori Friedman and Ashley Keefner, "Knowledge central: A central role for knowledge attributions in social evaluations," *Quarterly Journal of Experimental Psychology*, 70 (2006): 504–515; Turri, "Revisiting norms of assertion".

[22] See especially Edward Craig, *Knowledge and the state of nature* (Oxford University Press: Oxford, 1990); Michael Hannon, *What's the Point of Knowledge?: A Function-First Epistemology* (Oxford: Oxford University Press, 2019); Phillips et al., "Knowledge before belief," Section 6.

[23] Jonathan Phillips and Aaron Norby, "Factive theory of mind," *Mind & Language*, 36 (2021): 3–26.

[24] Or, at least, very close to it. See Adam M. Bricker, *Visuomotor noise and the non-factive analysis of knowledge* (Edinburgh: University of Edinburgh PhD Thesis, 2018); Wesley Buckwalter and John Turri, "Knowledge and truth: A Skeptical challenge," *Pacific Philosophical Quarterly*, 101 (2020): 93–101.

may be, as beliefs can very well be false. They have to be in the factive state of *knowing* about the topic.

Before moving on, it's worth noting that some have recently proposed that this factive dimension to theory of mind constitutes an entirely different kind of representational capacity, largely distinct from its non-factive counterparts.[25] Without the need to manage potentially competing perspectives (discussed below), this *factive mindreading* is hypothesized to be quicker and more automatic than non-factive mindreading, recruiting more efficient processes that demand far fewer cognitive resources. Westra and Nagel have even suggested that factive mindreading comprises the default mode for ordinary contexts like conversation, with non-factive mindreading only called upon in "special communicative contexts."[26] While I won't go quite this far, it is clear that the capacity to represent factive mental states like knowledge is especially central to human theory of mind capacities, playing a special role in flagging good informants, likely underwriting normative judgements about assertion and practical reasoning, and emerging earliest in both ontogeny and phylogeny.

Now that we have some sense of what theory of mind is, I want to talk a bit about how it might go wrong. While there are certainly many different ways in which we may fail to correctly represent the mental states of others—either by representing the wrong mental content, the wrong kind of state, or just missing a relevant state in its entirety—here I'll focus on three: epistemic egocentrism, contextual limitations on episodic simulation, and the outsized influence of stereotypes.

First, with the likely exception of factive states,[27] accurately representing the mental states of others will often require the management of representational content that doesn't match the content of one's own mental states. I can represent someone as believing that the earth is hollow despite myself not believing that. I can represent someone as wanting to eat meringue despite myself finding the stuff entirely unappealing. And I can represent someone as being deathly afraid of sharks despite myself fearing nothing piscine. This capacity to represent others as believing and desiring and fearing what we don't is central to the practical utility of theory of mind, which would be far more limited if we could only represent others as occupying mental states consistent with the states that we ourselves occupy. And central to this capacity is a neurocognitive process known as *self-perspective inhibition*. Roughly, when our own mental states are inconsistent with the content of others' mental states, we must inhibit that self-perspective content in order to represent the content of others' perspectives.[28] To represent you as believing that the earth is hollow, for example, I must inhibit my own belief that the earth is a solid cube.

[25] Jonathan Phillips, Wesley Buckwalter, Fiery Cushman, Ori Friedman, Alia Martin, John Turri, Laurie Santos and Joshua Knobe, "Actual knowledge," *The Behavioral and Brain Sciences*, 44 (2021): 177; Westra and Nagel, "Mindreading in conversation".

[26] Westra and Nagel, "Mindreading in conversation," 1.

[27] Adam M. Bricker, "The neural and cognitive mechanisms of knowledge attribution: An EEG study", *Cognition*, 203 (2020): 104412; Westra and Nagel, "Mindreading in conversation".

[28] Dana Samson, Ian A. Apperly, Umalini Kathirgamanathan and Glyn W. Humphreys, "Seeing it my way: a case of a selective deficit in inhibiting self-perspective", *Brain*, 128 (2005): 1102–1111; Lisette van der Meer, Nynke A. Groenewold, Willem A. Nolen, Marieke Pijnenborg and Aandré Aleman, "Inhibit yourself and understand the other: Neural basis of distinct processes underlying Theory of Mind", *NeuroImage*, 56 (2011): 2364–2374; Charlotte E. Hartwright, Ian A. Apperly and Peter C. Hansen, "The special case of self-perspective inhibition in mental, but not non-mental, representation", *Neuropsychologia*, 67 (2015): 183–192.

Crucially however, this inhibition doesn't always happen like it should. In a phenomenon known as *epistemic egocentrism*,[29] self-perspective information will often interfere with the accurate representation of others' inconsistent mental states.[30] As summarized by Nagel, "we overestimate the extent to which [others] share our beliefs, attitudes and concerns, even in the face of feedback to the contrary, and are surprisingly unaware of the extent to which we do this."[31] This is often referred to as *the curse of knowledge,*[32] as it's particularly easy to conceptualize in cases where one's knowledge about the world interferes with the ability to successfully represent another's false belief. Say, for example, I know that the coffee in my breakroom at work has been moved from behind the coffee maker to inside a cupboard, but, as I'm also aware, my colleagues do not know this yet. In such a case I may, erroneously, expect them to immediately go to the cupboard for the coffee, noticing a flash of surprise in myself when they first look behind the coffee maker. However, egocentrism certainly isn't limited to this kind of knowledge-ignorance asymmetry. I could just as easily make the mistake if, unbeknownst to me, after the coffee was moved to the cupboard, someone else moved it to the fridge. Now, it's important to be clear that egocentrism is not a certainty in any such cases. We can, and often do, successfully represent the false beliefs of others. But it is nonetheless a risk inherent in the practice of thinking about minds with content very different than ours, which plays out in empirically observable patterns of error.[33]

Moving on, we encounter a second, more general limitation on our ability to represent S's mental states when we don't have immediate access to the kinds of perceptual cues that underlie much of our theory of mind capacities, such as facial expressions, gaze direction, or vocal dynamics. If, for example, we're imagining what someone *would* think in a possible future scenario, we can't rely on any proximate visual or auditory information. We instead need to imagine what that person would think or desire or know through the process of *episodic simulation*, roughly projecting oneself into the possible episode and imagining how it might proceed.[34] Note that we've been doing a version of this in discussing counterfactual examples of mindreading throughout this section. Crucially however, empirical evidence suggests that how vividly we can imagine such episodes will influence how effectively we make inferences about the mental states of agents in these episodes.[35] For example, in one study from Gaesser et al.,[36] participants were asked to imagine simple, everyday episodes of helping others. These included helping someone who was locked out of their house or whose

[29] Don't confuse this with egocentric ignorance. They're two different things.

[30] For overviews, see Edward B. Royzman, Kimberly Wright Cassidy and Jonathan Baron, J. "'I Know, You Know': Epistemic Egocentrism in Children and Adults," *Review of General Psychology*, 7 (2003): 38–65; Jennifer Nagel, "Knowledge ascriptions and the psychological consequences of thinking about error," *The Philosophical Quarterly*, 60 (2010): 286–306, Section 4.

[31] Nagel, "Knowledge ascriptions and the psychological consequences of thinking about error," 302.

[32]See Susan A. J. Birch and Paul Bloom, "The Curse of Knowledge in Reasoning about False Beliefs," *Psychological Science*, 18 (2007): 382–386; Susan A. J. Birch, Patricia E. Brosseau-Liard, Taeh Haddock, and Siba E. Ghrear, "A 'curse of knowledge' in the absence of knowledge? People misattribute fluency when judging how common knowledge is among their peers," *Cognition*, 166 (2017): 447–458.

[33] See Birch and Bloom, "The Curse of Knowledge in Reasoning about False Beliefs"; Birch et al., "A 'curse of knowledge' in the absence of knowledge?".

[34] See Cristina M. Atance and Daniela K. O'Neill, "Episodic future thinking," *Trends in Cognitive Sciences*, 5 (2001): 533–539.

[35] Brendan Gaesser, Haley D. DiBiase and Elizabeth A. Kensinger, "A role for affect in the link between episodic simulation and prosociality," *Memory*, 25 (2017): 1052–1062; Brendan Gaesser, Kerri Keeler and Liane Young, "Moral imagination: Facilitating prosocial decision-making through scene imagery and theory of mind," Cognition, 171 (2018): 180–193.

[36] Gaesser, Keeler, and Young, "Moral imagination: Facilitating prosocial decision-making through scene imagery and theory of mind".

dog had gone missing. Participants imagined these episodes both in locations they were familiar with (i.e., somewhere they had been before and could easily imagine) and locations they were unfamiliar with (i.e., somewhere they hadn't been before). Most importantly for our purposes, participants reported significantly higher degrees of mentalizing—specifically considering the other person's thoughts and feelings in simulated episodes—when imagining helping someone in a familiar location.[37] These findings indicate that the extent to which we compute mental state representations during episodic simulation is modulated by our familiarity with the constituents of those simulations.

Gaesser summarizes this idea in the "episodic mindreading hypothesis," which posits that "the ability to remember and imagine specific episodes guides how targets' mental states within those episodes are perceived and, in some cases, what mental states are attributed. The details (e.g., location, objects, and agents) a target is retrieved with and bound to in an episode will constrain which mental states are assigned to the target."[38] To be clear, as with egocentrism, this isn't to say that we are incapable of successfully representing others' mental states when simulating unfamiliar conditions, only that our capacity to do so is more limited. Moreover, notice that in the case of episodic mindreading, the primary risk isn't that difficulty imagining an episode will result in mistaken mental state representations, but rather a kind of damping effect in which these representations simply won't be formed in the first place. As details in episodes become less familiar, our theory of mind systems just become less engaged.

Finally, I want to quickly call attention to the role of stereotypes in mental state attribution, a second way in which unfamiliarity with the constituents of an episode, particularly the agent herself, can impact our mindreading capacities.[39] It is likely that when evaluating the mental states of others, particularly those unfamiliar to us, stereotyping[40] might often play an outsized role.[41] As suggested by Westra:

> [O]ften, we interact with complete strangers, about whom we know nothing. In these cases, we may instead fall back on stereotypes about the target's social group membership. And this is a point where pernicious social biases can enter into the mindreading process distorting our interpretations of the social world.[42]

It's intuitive to understand why stereotypes might contribute to our representations of what unfamiliar agents think and know. Without other cues or information readily available, stereotypes—rapidly computed and easily accessible[43]—can quickly fill in the information vacuum. However, this brings with it a considerable risk for incomplete, distorted, or otherwise mistaken mental state representation. One example offered by Westra centers

[37] For summary, see Brendan Gaesser, "Episodic mindreading: Mentalizing guided by scene construction of imagined and remembered events," *Cognition*, 203 (2020): 104325, Section 5.3.

[38] Gaesser, "Episodic mindreading: Mentalizing guided by scene construction of imagined and remembered events," 4.

[39] I'd like to thank an anonymous reviewer for raising this point, which will return in Section 3.

[40] Here, stereotypes will be understood in the familiar sense of "conceptual attributes associated with a group and its members (often through over-generalization), which may refer to trait or circumstantial characteristics", David M. Amodio, "The neuroscience of prejudice and stereotyping," *Nature Reviews. Neuroscience,* 15 (2014): 670.

[41] Daniel R. Ames, Elke U. Weber, and Xi Zou, "Mind-reading in strategic interaction: The impact of perceived similarity on projection and stereotyping," *Organizational Behavior and Human Decision Processes*, 117 (2012): 96–110; Westra, "Stereotypes, theory of mind, and the action–prediction hierarchy".

[42] Westra, "Stereotypes, theory of mind, and the action–prediction hierarchy," 2822.

[43] Westra, "Stereotypes, theory of mind, and the action–prediction hierarchy," 2825.

around ageist stereotypes about memory: Citing studies that report that the memories and eyewitness testimony of both young children[44] and the elderly[45] are viewed as less reliable than other age groups, Westra summarizes, "We are much more likely to attribute false beliefs to both very young and old individuals than to other adults."[46] This reflects a more general pattern of the similarity or dissimilarity of an agent to us affecting how we represent their mental states,[47] highlighting the broad limitations that accompany mentalizing about the unfamiliar.

There is of course far more that could be said, not just about stereotyping but all the other aspects of mental state representation discussed here. We don't use theory of mind for just three things, and there are certainly more than three ways it might go wrong. Nevertheless, I do think that we have more than enough background in place to talk about how theory of mind is used in historiography. Let's do it, then.

3. Theory of mind in historiography

To start, I want to note that it isn't precisely clear how theory of mind is used in historiography, and I certainly won't venture to provide a systematic or definitive answer here. Instead, for now I only want to explore whether mental state representations, as used in historiography, might constitute a lower grade of evidence than they do in more familiar, everyday contexts. In doing so, we'll make two key observations: First, mental state representations are used in historiography to fulfill many of the same roles as in other applications, particularly action explanation but also normative evaluation. Second, however, the kinds of mentalizing engaged in historiography display textbook conditions conducive to epistemic egocentrism, diminished episodic simulation, and increased stereotyping. This raises questions about the evidentiary role of theory of mind in historiography, questions which require empirical techniques from cognitive science in order to sufficiently address. Note also that here I'll primarily focus on representations of the mental states of past agents as they feature in historical arguments and narratives, devoting less attention to the mental states of the authors of written historical sources.

In the previous section, we noted that one of the most basic roles of theory of mind is the explanation and prediction of others' behaviors and actions. The use of mental state representation in this role is readily observed in historiography. While I don't know that there's much need for action *prediction* when talking about the past, at least in the strict sense, the mental states of historical agents are frequently used to explain their actions. Consider the following examples from McPherson, in which he uses representations of fear states in a simple causal explanation of the secession of the US South at the beginning of the American Civil War:

[44] Gail S. Goodman, Jonathan M. Golding and Marshall M. Haith, "Jurors' Reactions to Child Witnesses," *Journal of Social Issues,* 40 (1984): 139–156; Gail S. Goodman, Jonathan M. Golding, Vicki S. Helgeson, Marshall M. Haith and Joseph Michelli, "When a Child Takes the Stand: Jurors' Perceptions of Children's Eyewitness Testimony," *Law and Human Behavior*, 11 (1987): 27–40.

[45] Katrin Mueller-Johnson, Michael P. Toglia, Charlotte D. Sweeney and Stephen J. Ceci, "The perceived credibility of older adults as witnesses and its relation to ageism," *Behavioral Sciences & the Law*, 25 (2007): 355–375.

[46] Westra, "Stereotypes, theory of mind, and the action–prediction hierarchy," 2824,

[47] See Brandon M. Woo and Jason P. Mitchell, "Simulation: A strategy for mindreading similar but not dissimilar others?", Journal of Experimental Social Psychology, 90 (2020): 104000.

> Eventually the expansion of free territory would make freedom the wave of the future, placing slavery "in course of ultimate extinction," as Lincoln phrased it. That was just what Southerners **feared**.[48]

> The cause of secession was one specific thing: the Southern response to the election of a president and party they **feared** as a threat to slavery.[49]

Structurally, this is no different than how we might use a fear representation in the causal explanation of actions in the present day. The *New York Times,* for example, makes ample use of this in their headlines—such as, "Fearing Backlash, G.O.P. is Quiet on Abortion"[50] and "Baltics, Fearing Russian Assault, Demand Tougher Stance from West."[51] As an aside, note that, as in the *Times* headlines, McPherson attributes a fear state not to a single individual, but instead a group agent, *the American South*. This, too, is perfectly in keeping with the ordinary practice of attributing mental states to groups.[52]

For another example, consider how Crone uses mental state representations to complement behavioral evidence in explaining why the French and British Empires of the nineteenth and twentieth centuries experienced far more secessionist sentiment than previous Arab empires in the same regions.[53] She begins with non-mentalized descriptions of behavior:

> The Europeans did recruit native administrators and soldiers but, unlike the Arabs, they did not have to use them in the top positions, let alone in the metropole itself, because they could keep sending new men from France and Britain for such posts.[54]

Before offering mental states that help explain and give context to this behavior:

> The French and British were not always **sure** that they really **wanted** an empire, since formal control was expensive and not always necessary for purposes of securing markets and raw materials; and they did not always **envisage** those colonies in which they did not settle as permanent possessions.[55]

Both behavioral and mental evidence are then used to support Crone's thesis about the exclusionary structure of European imperialism fomenting revolutionary nationalism:

> In short, even if the Europeans had expanded in Asia as bearers of churches rather than nations the conquered peoples could not have penetrated their ranks. Accordingly, the history of the French and British empires abounds in examples of secession by acculturated natives:

[48] James McPherson, *This mighty scourge: perspectives on the Civil War* (Oxford: Oxford University Press, 2007), 15; emphasis added.
[49] McPherson, *This mighty scourge*, 17; emphasis added.
[50] Jonathan Weisman, "Fearing Backlash, G.O.P. Is Quiet on Abortion," *New York Times*, May 7, 2022.
[51] Lara Jakes, "Baltics, Fearing Russian Assault, Demand Tougher Stance from West," *New York Times*, March 8, 2022.
[52] See Juan Manuel Contreras, Jessica Schirmer, Mahzarin R. Banaji, and Jason P. Mitchell, "Common Brain Regions with Distinct Patterns of Neural Responses during Mentalizing about Groups and Individuals," *Journal of Cognitive Neuroscience*, 25 (2013): 1406–1417; Adrianna Jenkins, David Dodell-Feder, Rebecca Saxe, and Joshua Knobe, "The neural bases of directed and spontaneous mental state attributions to group agents," *PloS One*, 9 (2014): e105341.
[53] Patricia Crone, *The nativist prophets of early Islamic Iran: rural revolt and local zoroastrianism* (Cambridge, UK: Cambridge University Press, 2012), chapter 8.
[54] Crone, *The nativist prophets of early Islamic Iran*, 173.
[55] Crone, *The nativist prophets of early Islamic Iran*, 173; emphasis added.

> they walked out as members of separatist churches, as leaders of nativist revolts, and above all as modern nationalists.[56]

Here too, these mental state representations function in much the same way they would in ordinary contexts, explaining the behaviors of the actors who hold them. I imagine that none of this should be particularly controversial, so let's move on.

Beyond explaining the actions of others, we also noted in the previous section that mental state representations are frequently used in the normative evaluation of actions and behaviors—not why someone engaged in some action, but whether they *should have*. Although perhaps to a lesser extent, this, too, is also readily observed in historiography.

Consider, for example, Farrell's discussion of Nixon's involvement in the cover-up of the Watergate break-in:

> Later on, when the cover-up collapsed, Senator Howard Baker, who had a gift for pithy encapsulation, would reduce the issue to: "What did the president **know**, and when did he **know** it?" The answer is that Nixon **knew** it all, and he **knew** it all along.[57]

Along with Nixon's knowledge that these political cover-ups are generally doomed to fail:

> Nixon **knew**, and reminded his aides until they grew tired of hearing it, that it was the cover-up, not the crime, that had brought down Alger Hiss. The traitor had gone to prison for perjury, not espionage. And so, Nixon said, Liddy and the burglars could not be protected. "The truth, you always figure, may come out," Nixon told Haldeman.[58]

With these two knowledge states, Farrell doesn't just explain Nixon's actions, but provides a normative appraisal of them. Nixon and his staff *should not* have engaged in any cover-up, not only because corruption is itself wrong, but because they knew better than to reasonably expect that a cover-up would work:

> Victory seemed assured. The cover-up, in retrospect, was ruinous overkill. Even had Nixon lost in November, he would have at least departed the White House in dignity, his foreign policy achievements intact and his own hands clean. Instead, Nixon's worst instincts, and those of his aides, betrayed him.[59]

Again, this is structurally no different than when we might, in everyday contexts, use an agent's knowledge in the normative evaluation of her actions.

Moving on, I'm less certain on the degree to which knowledge representations are used in historiography to flag good informants. As a purely anecdotal matter, it does seem more difficult to identify clear-cut instances of explicit knowledge attribution from egocentric ignorance. This, of course, doesn't rule out the possibility that historians might utilize *implicit* egocentrically ignorant knowledge representations, particularly when evaluating historical sources. The authors of testimonial evidence—the written and oral records that historians rely on for much of their work—are plausible candidates for being represented as knowers, at least in many instances. However, as a more conceptual point, we can note that the way in

[56] Crone, *The nativist prophets of early Islamic Iran*, 173.
[57] John Farrell, *Richard Nixon: The Life* (New York: Doubleday, 2017), 480; emphasis added.
[58] Farrell, *Richard Nixon*, 475; emphasis added.
[59] Farrell, *Richard Nixon*, 474.

which historians engage with sources is categorically different than what we observe in ordinary factive mindreading. Rather than operating on a default assumption that past agents' representations match one's own representation of reality, historical sources are subject to systematic, reflective scrutiny. This is especially salient in the tradition of source criticism, which prompts historians to ask questions like, "Was the author of the text in a position to know what he reported? Did he intend an accurate report? Are his interpretations 'reliable'?"[60] This is unambiguously non-factive mindreading. Although it may ultimately result in a knowledge representation, it certainly doesn't do so by default. In a sort of inversion of the picture provided by the factive mindreading hypothesis, non-factive mindreading appears to be the default, with the move to factive only coming upon finding sufficient reason to do so.

To summarize, while there are likely asymmetries in whether and how factive theory of mind is used to flag good informants, we have good reason to think that theory of mind capacities are deployed by historians to fulfil many of the same roles as they are in everyday contexts. Normative evaluations and, in particular, action explanations in historiography both rely on mental state representations in familiar ways. I now want to raise the question of the quality of these representations within the context of historiography. Put briefly, there is reason to suspect that they constitute a poorer grade of evidence in reasoning about causal explanation and normative evaluation than they might in everyday contexts. The conditions under which we think about past minds are particularly ripe for epistemic egocentrism, impoverished episodic simulation, and stereotype-driven distortions.

Let's start with epistemic egocentrism. As discussed in the previous section, this "curse of knowledge" frequently occurs when we have to evaluate the mental states of agents that know less than we do, don't have access to the same information, or otherwise hold mental states inconsistent with our own. In virtue of some fairly immutable features of time, this epistemic imbalance is particularly pronounced in historiography. We know how history goes in a way that past agents simply cannot. The American South of 1860 doesn't know that war will break out the following year, that the South would lose that war, or that slavery is soon to be abolished. French and British imperialists around the turn of the twentieth century don't know that their Empires are soon to break apart. Nixon in 1972 doesn't know that attempts to cover up the Watergate break-in will ultimately be exposed in dramatic fashion, leading to his resignation of the presidency and cementing his reputation as one of the most reviled political leaders in US history. But the historian knows all these things. She is in a position of considerable epistemic advantage, the exact conditions that generate egocentrism. This raises the real possibility that, in many instances, historians' representations of past mental states may be influenced by their own privileged epistemic standing. It's easier, for example, to attribute a fear state to the American South knowing that these fears come to pass; it's easier to think that the French and British didn't envisage their colonies as permanent when one knows of their ultimate impermanence; and it's easier to say that Nixon knew that cover-ups, as a rule, inevitably fail when one knows that Nixon's infamous attempt was itself doomed to failure, subsequently constituting American culture's paradigm case of the failed cover-up. Whether egocentrism plays a causal role in these kinds of cases is much more difficult to answer. However, at a minimum, all the hallmarks are clearly there.

[60] Martha Howell and Walter Prevenier, *From reliable sources: an introduction to historical methods* (Ithica, NY: Cornell University Press, 2001), 60.

We find a similar situation for episodic simulation. When thinking about past minds, we cannot rely on any of the proximate cues that underwrite much of our ability to reason about the mental states of others. There is no ready access to the gaze direction, facial expressions, or prosody of historical figures. As with reasoning about mental states in possible futures, thinking about the minds of past agents will require a good deal of episodic simulation, imagining the past in order to mentalize about it. However, here our theory of mind capacities encounter another fundamental limitation. Broadly speaking, historical episodes are by their nature unfamiliar. As they move farther away from the present, they will involve agents, locations, and objects—as well as beliefs, norms, and social practices—that are increasingly difficult to imagine with the vividness required of successful mental state evaluation: the Nixon Whitehouse, Haldeman and Liddy; Northern Africa circa 1900, French and British colonial power; the Antebellum South, the abolitionists of the nascent Republican party, slavery as an American social institution. To be clear, it's not that such historical entities are intrinsically unimaginable in the present, but rather that the bar for engaging theory of mind in the first place is likely quite high. Recall that even an unfamiliar location for an imagined everyday episode, like helping someone locked out of their house[61], can damp levels of mentalizing. Accordingly, there is a real risk that mentalizing for historical episodes—orders of magnitude more unfamiliar—is damped to a non-trivial degree, resulting not in less accurate mental state attribution, but simply less.

Finally, we can quickly notice that similar reasoning extends to the potential for increased reliance on stereotypes in reasoning about past minds. The agents of the past are particularly unfamiliar to us, boosting the chances that stereotypes will contribute to our judgements about their mental states. This may be especially likely in the case of group agents—for example, "Europeans" or "the South"—for which the cultural availability of stereotypes are heightened. But notice that here, unlike with episodic simulation, the risk *is* that we may attribute incorrect mental states to past agents. Rather than simply damping theory of mind, overreliance on stereotypes can result in distorted patterns of mental state attribution, like the ageist bias towards the attribution of false belief discussed above.

Unfortunately, here we can't go much further than speculation and vague generalities. We can at least identify a legitimate epistemic concern. Historiography fulfils textbook conditions for cognitive limitations in mental state representation. Nevertheless, this doesn't itself entail that these limitations are inevitably insurmountable, or even problematic. It could well be that historians take enough deliberate care when thinking about past minds to inoculate themselves against egocentrism, or that years of dedicated study provides sufficient familiarity to overcome the inherent challenge of vividly imagining the past. The point I want to make is not that egocentrism, damped episodic simulation, and stereotype-driven distortions are guaranteed to be a major problem for historiography, but rather that the philosophy of historiography is not itself equipped to answer whether they are. This is fundamentally an empirical problem, which can only be addressed through empirical methods. To close this paper, I'll say a word on what such an empirical project might look like. But first, I want to explore how these considerations could have deeper philosophical implications, particularly for approaches that place theory of mind as central to our understanding of history.

[61] Gaesser et al., "Moral imagination".

4. Collingwood: A brief case study

Let's assume, for a moment, that theory of mind capacities are indeed degraded in historiographical applications, with pervasive epistemic egocentrism, severely impoverished episodic simulation, and/or stereotype-driven distortions resulting in significantly less reliable mental state representations than we'd expect in everyday contexts. What would the deeper philosophical significance of this be? Would it threaten the possibility of historical knowledge? Is theory of mind reasoning best understood as merely a contingent fact about the practice of historians, or a necessary step in understanding the past? Answering these questions in any comprehensive way is far beyond my ability to deliver here, as this will largely depend on additional philosophical assumptions about the nature of history and historiography. Will it mean the same thing for Kuukkanen's postnarrativism[62] as its narrativist predecessors? Probably not. Observe also that the use of theory of mind in narrative explanation by historians doesn't itself mean this practice is necessary.[63] In many cases, working out philosophical from merely methodological consequences will be anything but trivial. Accordingly, rather than attempting a systematic evaluation of all such consequences here, I want to focus on one particular account where these consequences are more straightforward—Collingwood's *The Idea of History.*[64] In what may well be the worst-case scenario given compromised thinking about past minds, Collingwood puts forward a view on which mental state representation is central not just to the contingent practice of historians, but the philosophical foundations of history itself.

Collingwood's account emerges from much the same kind of epistemological question that has motivated us here, "How, or on what conditions, can the historian know the past?" To gain knowledge of the past, Collingwood writes, "the historian must re-enact the past in his own mind."[65] This suggestion is itself noteworthy because of the striking similarity to the episodic simulation discussed above. Collingwood is naturally understood as describing something very much along the lines of projecting oneself into the past, armed with some "documents or relics" and tasked with imagining "what the past was which has left these relics behind."[66] More significant for our purposes, however, is the *kind* of reenactment that Collingwood describes. It is, at its core, a process of mental state representation, on which the historian reenacts the thoughts and experiences of past agents.[67] This is nicely illustrated with Collingwood's example of a historian reading the *Codex Theodosianus* (a set of Imperial Roman laws):

> Suppose, for example, he is reading the Theodosian Code, and has before him a certain edict of an emperor. Merely reading the words and being able to translate them does not amount to knowing their historical significance. In order to do that he must envisage the situation with which the emperor was trying to deal, and he must envisage it as that emperor envisaged it. Then he must see for himself, just as if the emperor's situation were his own, how such a situation might be dealt with; he must see the possible alternatives, and the reasons for choosing one rather than another; and thus he must go through the process which the emperor

[62] Jouni-Matti Kuukkanen, *Postnarrativist philosophy of historiography* (New York: Palgrave Macmillan, 2015).

[63] See Mariana Imaz-Sheinbaum and Paul A. Roth, "The end of histories? Review essay of Alexander Rosenberg's how history gets things wrong: The neuroscience of our addiction to stories," *Journal of the Philosophy of History*, 15 (2020): 240–248.

[64] Robin G. Collingwood, *The idea of history* (Oxford: Clarendon Press, 1946). 1966 Reprint.

[65] Collingwood, *The idea of history*, 282.

[66] Collingwood, *The idea of history*, 283.

[67] Collingwood, *The idea of history*, 283.

> went through in deciding on this particular course. **Thus he is re-enacting in his own mind the experience of the emperor; and only in so far as he does this has he any historical knowledge**, as distinct from a merely philological knowledge, of the meaning of the edict.[68]

Elsewhere, Collingwood offers a series of other examples that illustrate this foundational requirement of historical knowledge, including the following endorsement of the possibility of a history of economic activity:

> A man who builds a factory or starts a bank is **acting on a purpose which we can understand;** so are the men who accept wages from him, buy his goods or his shares, or make deposits and withdrawals. If we are told that there was a strike at the factory or a run on the bank, we can **reconstruct in our own minds the purposes of the people whose collective action took those forms**.[69]

All this is about as unambiguously (and strongly!) mentalistic as you can get. Mindreading is not merely a central part of what historians do, but it is only through mindreading that historical knowledge is possible. "All historical thinking," writes Collingwood, "is thinking about the act of thinking."[70] Notice that for Collingwood, thinking is explicitly understood in the familiar mentalistic sense, "a certain form of experience or mental activity" that is "not merely immediate, and therefore is not carried away by the flow of consciousness."[71] Unsurprisingly, then, Collingwood defends in detail the possibility of accurately understanding the minds of past agents,[72] recognizing that his account will only be as good as our ability to represent past mental states. To the contemporary eye, perhaps the most recognizable objection Collingwood considers ultimately reduces (at least as Collingwood argues) to a familiar expression of solipsism, "the doctrine that my mind is the only one that exists."[73] Perhaps more noteworthy, however, is *how* Collingwood characterizes knowledge about other mental states, writing: "To know someone else's activity of thinking is possible only on the assumption that this same activity can be re-enacted in one's own mind."[74] In using this language of reenactment, Collingwood's account almost seems to anticipate modern simulation theory, which views theory of mind as functioning through simulating the mental states of others using the cognitive processes that also support those mental states in the mind of the evaluator.[75] All told, we find a remarkably clear-cut instance in which theory of mind is placed at the very center of our knowledge of the past.

Unfortunately, however, Collingwood is of course unable to anticipate objections that arise from modern theory of mind research. While it's certainly not fair to expect Collingwood (d. 1943) to have imagined potential cognitive limitations of thinking about past minds, it is clear that any such limitations would present a considerable challenge for his account. To whatever extent that egocentrism, limited episodic simulation, or stereotype-driven distortions might degrade our capacity to think about the thoughts of past agents, they will, on Collingwood's view, undermine our capacity to have historical knowledge. On the assumption that such

[68] Collingwood, *The idea of history*, 283; emphasis added.
[69] Collingwood, *The idea of history*, 310; emphasis added.
[70] Collingwood, *The idea of history*, 307.
[71] Collingwood, *The idea of history*, 306.
[72] Collingwood, *The idea of history*, 283–302.
[73] Collingwood, *The idea of history*, 288.
[74] Collingwood, *The idea of history*, 288.
[75] See Robert M. Gordon, "Folk Psychology as Simulation" *Mind & Language*, 1 (1986), 158–171; Alvin I. Goldman, *Simulating minds: the philosophy, psychology, and neuroscience of mindreading* (Oxford: Oxford University Press, 2006).

cognitive effects ultimately prove detrimental, this would then raise the specter of a rather severe variety of skepticism of knowledge about the past. Much, if not most, historical knowledge would be under threat. As is the case with any skeptical problem, I'm not sure that this would prove insurmountable. Nevertheless, it would present a substantial challenge for the view.

Of course, all of this is significant only if the potential cognitive limitations I've described above do indeed prove to be actual, empirically verifiable impediments to our ability to represent past mental states. To close out this paper, then, I want to sketch what the empirical investigation of these matters might look like in practice. If there is good reason to suspect that theory of mind is limited in historiographical contexts, how are we to test whether our reasonable suspicions are correct?

5. Conclusion: One possible future philosophy of historiography

At the center of this paper is a thesis about the representations of past mental states—the quality of evidence, constituted by mental state representations within the context of historiography, is undermined by cognitive limitations of theory of mind.[76] Here we've made a number of observations about this thesis: It falls squarely under the purview of the philosophy of historiography; the philosophy of historiography is not presently equipped to address the thesis; the thesis is at least theoretically plausible; the thesis, were it correct, would have identifiable philosophical consequences; so, all told, we have sufficient reason to submit the thesis to further scrutiny. But all this raises the question of just how exactly we are to investigate such a claim. Returning to the ideas set out at the beginning of the paper, I'd propose what we need here is a research program that merges the methods of experimental cognitive science, with an assist from digital humanities, into the questions and subject matter of the philosophy of historiography—an integrated cognitive science of historiographical representation.

There are at least three different components identifiable for such a program, each with its own set of methodological demands and requirements for technical expertise. First, there is the preliminary (but no less significant) task of describing how, exactly, mental state representations are used in historiography. For example: Is there a significant difference between explanatory and normative roles? What kinds of mental states are most frequently represented? And do certain kinds of histories, like biographies, use mental state representations differently than others? This descriptive project is likely a task for the digital humanities, which can quantitatively analyze large volumes of written work to gain a more complete understanding of how historians use mental state representations.

A second task, which relates most closely to the questions raised in this paper, is understanding to what extent cognitive limitations might contribute to decreased reliability in mental state representation when thinking about past minds. To some extent, this too could be a job for digital humanities. Certain textual patterns might provide evidence that these limitations drive judgements about past mental states. A very high proportion of representations with content that matches the historian's own mental states may suggest

[76] To be clear, this isn't to suggest anything like a wholescale skepticism about our representation of past mental states. It could still be, and very likely is, that historians often correctly evaluate the minds of past agents. The thesis explored in this paper is that such evaluations may not constitute the same high-quality evidence that they do in everyday, present conditions.

egocentrism, and an inverse correlation between mentalizing and distance from the present (or some other proxy for unfamiliarity) may indicate simulation-driven effects. However, of primary importance here are the methods of experimental cognitive science—running experiments in which participants compute actual mental state representations (e.g., reading vignettes that describe an agent in some scenario) on the basis of both everyday information and the kind of information available to historians. This could then provide evidence directly relevant to the question of whether historiographical contexts are particularly conducive to the diminished reliability of theory of mind capacities.

Finally, there is the task for traditional philosophy of history and historiography: understanding the philosophical implications of the empirical findings provided by both cognitive science and textual analysis. What philosophical accounts are impacted by limitations in thinking about past minds, and to what extent? Is this isolated to Collingwood, or does it generalize? Are these cognitive limitations merely contingent facts about the kinds of evidence historians happen to use, or do they correspond with necessary features of historical knowledge?

Admittedly, none of this will be particularly easy to execute. It would require a significant degree of cross-discipline collaboration between cognitive science, the digital humanities, and traditional philosophy of history/historiography. Cognitive scientists would need the input of philosophers of history to ensure their experiments are appropriately valid, telling us something that meaningfully extends to the cognition of actual historians when they do actual historiography. Likewise, philosophers of history would need the input of cognitive scientists to ensure that their experimental findings are appropriately interpreted when unpacking philosophical implications. Digital humanities specialists would need the input of both cognitive scientists and philosophers of history to separate genuine mental state representations from mental state terms used in historiographically uninteresting ways. And so on. However, despite this high bar to entry, the research program I've advocated for here is one of considerable promise. It has the potential to provide important new insights into how one of the most foundational elements of human social cognition functions within one of the most important human intellectual endeavors. And, moreover, it's likely that other disciplines in the social sciences that make use of theory of mind for unfamiliar or distant agents—for example, sociology, political science, anthropology, and literary studies—may encounter similar limitations, further highlighting the need for establishing methods to investigate how (and how well) theory of mind is deployed in such contexts.[77] Crucially, however, none of this is achievable without crossing a number of traditional disciplinary boundaries. While it may be quite the effort, it's my opinion that this effort would be well worth it.

[77] I'm thankful to both an anonymous reviewer and the audience at the 2022 workshop for raising this point.