Learning through Simulation

Sara Aronowitz

Program in Cognitive Science & University Center for Human Values

Princeton University

skaron@umich.edu

Tania Lombrozo

Department of Psychology

Princeton University

lombrozo@princeton.edu

Abstract

Mental simulation -- such as imagining tilting a glass to figure out the angle at which water would spill-- can be a way of coming to know the answer to an internally or externally posed query. Is this form of learning a species of inference or a form of observation? We argue that it is neither: learning through simulation is a genuinely distinct form of learning. On our account, simulation can support learning the answer to a query even when the basis for that answer is opaque to the learner. Moreover, through repeated simulation, the learner can reduce this opacity, supporting self-training and the acquisition of more accurate models of the world. Simulation is thus an essential part of the story of how creatures like us become effective learners and knowers

Introduction

Imagine that you have two cylindrical glasses of the same height, where one is wide and one is thin. Each glass is filled with water to the same height. If you slowly tip both glasses over at the same rate, which glass will spill water first? Or, will they both spill water when tipped to the very same angle?

This challenge – known as the "water-pouring problem" (Schwartz & Black, 1999) – is often answered incorrectly. But if people are invited to imagine holding out their hands (as if holding both glasses), to close their eyes, and to mentally rotate their hands until they think the water would spill, they often produce the correct response: that water from the wide glass will spill first. Call this activity of answering a query by imagining and then evaluating the output of one's imagination *mental simulation*.

The exercise of simulation just described looks like a *way of learning something new*. Construed in this way, it is a lot like learning through observation. But it's also natural to say that this kind of learning *doesn't rely on acquiring any new information*. Described like this, it is a lot like learning through inference.

How does learning through simulation support learning? And what, if anything, is epistemically distinctive about learning through simulation? Can learning through simulation be reduced to learning through observation or through inference?

This paper aims to make progress on these questions. We argue that mental simulation cannot be reduced to observation or inference, though it shares important similarities with both. On our account, mental simulation is distinctive in two ways. First, simulation is a method that can support learning some proposition *q* (the output of the simulation) even when the basis for that output is opaque to the learner. Second, through repeated simulation, the learner can reduce this opacity, supporting the acquisition of more accurate models of the world. This allows us to see why (and under what conditions) simulation is an essential part of learning from what we already know.

Our account is partially motivated by new developments in machine learning and artificial intelligence that make the topic of simulation timely. Some of today's most successful artificial agents (most famously the deep learning architecture AlphaGo, Silver et al., 2016) learn through simulation, training themselves over and over on simulated outcomes. Within reinforcement learning, comparing model-based and model-free systems has opened an avenue of research into the power and limits of an internal model (e.g., Atkeson & Santamaria, 1997; Hamrick, 2019). And even the most basic unsupervised clustering methods have challenged our assumptions about how much can be learned through a process which, like many simulations, is opaque to the person relying on it, and devoid of explicit knowledge.

These examples complement more familiar examples of mental simulation that have been the subject of prolonged interest and debate in both psychology and philosophy, with important work on the role of imagination (e.g., Kind & Kung (Ed.), 2016) and thought experiments (e.g., Brown & Fehige, 2017). As we use the term, however, simulation is a more expansive category than thought experiments: it includes cases of repeated ("long-run") simulation, as in fine-tuning athletic skills through motor imagery, as well as artificial simulations. Theorizing about this larger category brings into focus features of learning through simulation that have been under-emphasized in past work on the epistemic role of simulation in the mind and in science, in particular the long-run use of simulation and the connections between one-shot and long-run simulation.[1]

Of course, drawing such a wide net also has disadvantages. Why think this heterogenous category will have anything interesting in common? One reason to suspect there *is* an epistemic feature shared by most simulations is the similarity of the philosophical debates surrounding their use.

[1] A side-effect of this wider focus is that our account does not differentiate between simulations that are conscious or unconscious, internal or external, or propositional or non-propositional. While these are interesting differences, we aim to establish a similarity in the epistemic function of the process across these dichotomies.

Shannon Spaulding (2016) argues that mindreading[2] (simulating the thoughts of others) cannot be a way to learn new things about the world, but is a way of generating ideas to assist other cognitive faculties that are capable of producing knowledge through inference (or observation). Tamar Gendler (2004) advances the view that thought experiments are often "quasi-observational," such that accompanying visual imagery plays a critical epistemic role. Discussing computer simulation in physics, Eran Tal (2009) asserts that simulation can confirm hypotheses, but not by inferring from data to theory - instead, the simulations he discusses offer a way of inferring from theory to data. In each case, we see a remarkably similar dialectic, attempting to fit simulation into some epistemic role relative to inference and observation.

A second reason to draw a wide net in efforts to capture the epistemic role of simulation comes from the success of related projects: philosophers already take inference and observation to be categories that cover both individual cognitive exercises and public scientific acts – and so it seems apt to ask a similar question about simulation at the same level of generality. However, in order to keep our argument focused, we will build a theory of mental simulation around evidence from a single domain, motor simulation, and then argue that the account generalizes to scientific and artificial contexts.

By making progress in understanding the function of simulation, we'll also shed light on the function of the *capacity* to simulate. Why would this capacity be useful in the first place? Simulation is a fairly common mental activity, at least in humans, and yet it's not obvious that if a designer were to create a learning machine to solve the kinds of problems we solve, she would give the machine the capacity to simulate. Our account sheds light on why we are the kinds of learners that benefit from simulation.

We start in Section 1 with a case from machine learning that presents a functional role for simulation – self-training – that a successful account should capture. In Section 2, we'll clarify our assumptions and terminology concerning simulation, observation, and inference as ways of acquiring knowledge. Sections 3 and 4 explore ways in which simulation might be related to observation and inference. We conclude that while simulation has something in common with each, it cannot be reduced to a version of either. Sections 5-7 present our positive account: we explain what simulation lacks (Section 5), what it offers (Section 6), and how it works (Section 7), introducing the important idea of *representational extraction*. We conclude in Section 8 by discussing the ramifications of our proposal.

1. An Epistemic Function of Simulation: Self-training

In this section, we'll run through one example of a machine learning process that trains its own internal model through a kind of simulation. This case is not meant to be particularly unique, but rather an exemplar of a general feature that helps explain the success of this class of learning methods – and raises questions about how humans and other animals might build and use internal

[2] A long and complex debate exists over whether theory of mind (or mindreading) is a simulation, an inference, or involves both. We'll avoid taking a position in this debate, and we don't intend the discussion of inference here to necessarily track the sense of inference invoked in the "theory-theory" (e.g., Sellars, 1956).

models.

Our example is a model-based reinforcement learning method called "Hallucinated Replay" (Talvitie, 2014). In contrast to more standard reinforcement learning algorithms (such as Q-learning; Sutton & Barto, 1998), Hallucinated Replay uses two kinds of feedback to learn. The first kind of feedback is ordinary, external feedback: the model is updated when actual observations fail to conform to predictions. The second kind of feedback involves "hallucinated" rather than actual observations: if a predicted state (e.g., that was expected to lead to some reward) is not observed, the model will update expectations with respect to the *predicted* state as well as the *actual* state, even though the former was not actually observed.[3] Interestingly, the predicted ("hallucinated") states need not be possible states of the environment, so they would *never* be observed and subsequently updated using more standard forms of training. The hallucinated training, on the other hand, essentially trains the model to get back on track from these fictional states. This algorithm is thus training itself to predict well, even when it predicts based on incorrect imagined "feedback."

Hallucinated Replay is not globally superior to other model-based reinforcement learning algorithms. However, it has been shown empirically and analytically to succeed in contexts where comparable methods, trained on the same data (but without "hallucination"), tend to struggle (Talvitie, 2017, 2018). In particular, the model does better in environments that require agents to correct their model structure rather than merely tune parameters. Hallucinated Replay seems to help with this process because the model trains itself to self-correct from a wide range of inaccurate states, including those that would never be produced by the actual environment.

This example reflects a ubiquitous feature of machine learning algorithms: to succeed, they must be able to train themselves, and this training cannot depend on a representation of the reward structure of the environment that is already accurate.[4] In developing our own account of learning through simulation, we draw a key lesson: that simulation is an exercise performed using an internal model *that also trains and alters that model*. We refer to this feature as *self-training*, and a motivating consideration for our account of learning through simulation is that it accommodate this form of learning alongside more familiar cases of learning from isolated episodes of simulation, such as the water-pouring problem and most scientific thought experiments.

2. Clarifications

Before moving on to consider learning through observation and through inference, it's worth

[3] As a quick illustration, imagine the agent starts in state S_0, and predicts initially that act A_1 in S_0 will bring about S_1, and then act A_2 in S_1 will bring about S_r, a rewarding state. However, the agent is wrong: act A_1 in S_0 brings about S_{1*}, and act A_2 in S_{1*} brings about S_{r*}. A standard update would train the model at S_0 against the actual observed transition to S_{1*}, generate a prediction for the acts available at S_{1*}, and then evaluate that prediction against the next observation. In this way, only predictions from actual states get updated. Hallucinated Replay adds a second update, where taking A_1 in the "hallucinated" state S_1 is corrected as if it should have resulted in S_{r*} rather than the predicted S_r.

[4] See Hamrick (2019) for a review of simulation in reinforcement learning algorithms.

introducing some clarifications of our aims and terminology.

First, note that our aim is to put forward a new theory of an epistemic function of simulation, not to identify the *only* function of simulation. A further qualification is that while capturing this distinctive function of simulation will bring into focus differences between simulation and paradigmatic cases of observation and inference, there are other ways of inferring and observing that are more similar to, and perhaps intertwined with, simulation. Our reason for focusing on paradigmatic cases of observation and inference is that we have a good grasp on how such cases generate knowledge (even if the details are controversial). Assimilating simulation to observation or inference is thus attractive as a solution to the puzzle of how simulation can be a way to generate knowledge. Rather than drawing a bright line between simulation on the one hand, and observation and inference on the other, however, we instead hope that our final account will illuminate important similarities, and provide a foundation for future work that takes seriously the elements of simulation and imagination that are embedded in perception, induction, prediction, and other ways of observing or inferring.

We will compare three kinds of learning processes, defined as ways of coming to know the answer to an internally or externally posed query. These processes are learning through simulation, learning through observation, and learning through inference. We'll describe each process in terms of its inputs and output. The former could include inputs from the external world, from the thinker's internal representations, or some combination thereof. These inputs are utilized by some kind of cognitive process to yield the output: a piece of knowledge, knowledge that q, which we denote *K(q)*. Note that the sense of learning we are invoking here is factive: learning that q entails that q. We also assume that knowledge requires more than true belief, but the way that this additional commitment is spelled out will not matter for our story.[5]

To illustrate these processes in action, consider again the water-pouring problem. We assume that holding out your hands to simulate pouring, or creating a vivid mental image of so doing, constitute simulation. However, you could have solved this problem a different way: by finding two glasses fitting the description, filling them with water to the same level, and then rotating them to see which spills first. In this case, you did not simulate to arrive at an answer, but instead set up an experiment and then learned the result through observation.

Likewise, when presented with the water-pouring problem, you might have solved the problem using static images, like those in Figure 1, and an argument along the following lines.

[5] We'll discuss learning in terms of coming to know a proposition, but it's worth noting that on many accounts, there are separate and non-reducible ways of coming to know a skill or coming to know an object (in the sense of acquaintance). We suspect our arguments hold equally well for these other kinds of knowledge, but fleshing out how this would work is outside the scope of the present paper.

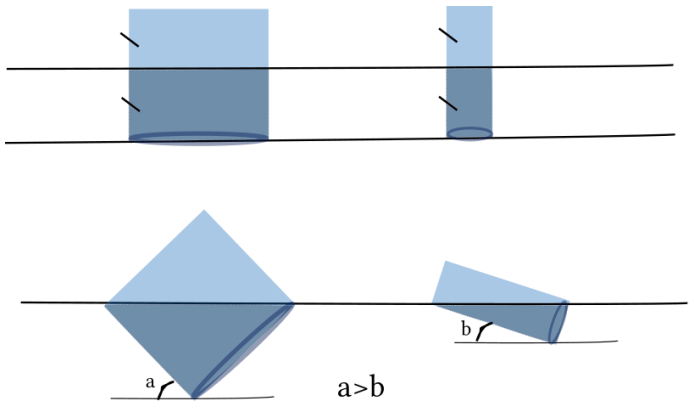

Figure 1. A representation of the water-pouring problem.

Say that the wide glass is just as tall as it is wide, and both glasses are filled up to exactly halfway. In that case, the liquid will spill out of the wide glass as soon as it is turned past a 45-degree angle, since the diagonal line from the top left to the bottom right of the glass divides it by volume into two equal halves, just like the original horizontal line. On the other hand, the thin glass, when divided by the same diagonal line, will have to be tipped much farther to keep the diagonal line parallel to the horizontal. This explanation shows the same result as the motor simulation, and even uses a visual device in the form of a figure, but intuitively it is not a simulation. As just described, this process of figuring out the answer is instead a kind of learning through inference.

With these paradigmatic cases of simulation, observations, and inference as reference points, we can move on to our arguments.

3. Simulation as Learning through Observation

One way to make sense of learning through simulation is to assimilate it to learning through observation. In this section, we first develop this picture, and then argue that it is inadequate.

We can start with a rough definition of learning through observation as a transition from the representation of some external piece of new evidence *e* to a conclusion *q*. *S*'s learning *q* from evidence *e* realizes a relation of external evidential support, such that *e* renders *q* true or likely to be true. In the terminology introduced in the previous section, this means that the input to learning includes new evidence, and that the cognitive process realizes the relation of external evidential support. This external evidential support relation is necessary to explain why being exposed to an *apple*, but forming a subjectively justified internal representation *as of an orange*, does not count as an observation. Whatever else one might require of learning from observation, it should be relatively uncontroversial that what it means to observe requires that the target of observation stand in some sort of external evidential relation to the internal representation it produces.

Simulation has at least a superficial similarity to observation. In our examples, the thinker sets up some conditions and then "observes" the outcome of her simulation. Pursuing this parallel, we can see both observation and simulation as answering the same query (e.g., "which cup will spill first?"), and as doing so in much the same way: through a transition from some (actual or simulated) observation *e* to some conclusion *q*.

In fact, some form of equivalence between actual and simulated observations is advocated or presupposed in empirical research on learning from simulated experience. As one example, Kappes and Morewedge (2016) argue that simulated experience can sometimes "substitute" for the corresponding real experience, because "mentally simulating an experience induces equivalent downstream psychological and behavioral effects as actually having the corresponding experience" (pp. 406-407). They review evidence from a range of cases, including the mental simulation of hypothetical observations, sequential procedures, or desired outcomes.

Another example comes from the literature on motor imagery, where the dominant paradigm is what is called the "functional equivalence" model, originally expressed as the idea that "imagery of movements has some functional effects on motor behaviour that are in some way equivalent to actual movements" (Johnson 1982, pp. 363). Jeannerod suggests that "it seems a logical consequence of this rehearsal of the corresponding brain structures, and specifically the motor structures, that the subsequent execution will be facilitated" (pp. 108). Of course, advocates of functional equivalence do not consider actual motor experience and motor simulation completely equivalent: for instance, motor training might help stretch out muscles, but surely motor simulation does not have that function. Rather, this equivalence is restricted to a *subset* of the functions of each activity, where this subset is taken to explain *how* motor simulation supports learning.

In contrast to this thesis, we'll argue that assimilating the functional role of simulation to that of observation misses one of the most important ways in which simulation supports learning. Like the two forms of feedback in Hallucinated Replay, actual and simulated observations play different functional roles when it comes to *self-training*, and understanding the role of simulation in self-training is crucial for understanding (e.g.) how motor simulation supports learning.

To see why observation and simulation come apart when it comes to their functional roles, it's useful to consider two time-points at which we might engage in simulation: at the beginning of learning (when the simulation mechanism is not yet trained), or at the end of learning (when the simulation mechanism is fully trained). Of course, real cases will almost always fall within these time-points; for our purposes, these idealized points are merely relative benchmarks.

To illustrate, consider a simulation of a golf swing and an actual exercise of a golf swing, which both have as their output a state with the content that the ball misses the target by a few inches to the right. A functional equivalence theorist might be tempted to say that each of these processes yields an "observation," and that these observations serve the same function. And in some sense this is right: both events could have the same immediate impact on the motor system, and both serve the function of allowing the simulator to generate a response to the question of where the ball will go. Yet the simulated swing in fact generates something distinct from actual observation: it generates (likely noisy) information about where the ball would go *on the simulator's current motor model.*[6] This point becomes important when we consider the beginning of learning – before

[6] In fact, even in the one shot case we can observe what look like systematic differences between an imagined motion and an actual motor execution; Walsh and Rosenbaum (2009) find that imagined motions and actual motions conform to different aims: for instance, in imagining a line connecting to an ellipse we tend to prefer connecting the line to the apex, whereas when drawing, we tend to prefer a shorter path.

the model is fully trained.

Consider a case where your motor model is biased to the left: you tend to predict that balls you hit will land farther to the left than they actually do, were you to replicate the imagined swing in practice. In this case, the actual swing gives you a baseline against which to judge your simulated swing: when they have different results, especially in this systematic way, you should start to ask yourself where this difference is coming from and conclude that you are not taking into account some critical factor in your imagination. The combined use of the motor exercise and motor simulation allows you to solve a problem: you can now adjust the motor model for better prediction in the future. But the reason why the two swings can work together in this case is that *each serves a different function* – neither two motor exercises nor two motor simulations could enable you to correct your model in the way described above. Actual swings are the most direct way to learn what's off about your simulation, and simulated swings are the most direct way to learn what your current knowledge base and motor model predict.

Thus, focusing on the beginning of learning illustrates why the functional equivalence claim breaks down. If the two processes were truly equivalent *for training*, then at a process level, observation and simulation would "calibrate" or self-train the model in just the same way. But the lesson from such cases is more general: even for highly trained experts, relying on the model and relying on feedback from the environment to learn a single output may have different costs and be reliable under different conditions, insofar as even perfect experts will still need to calibrate and tune when they face changes in their environments or in their own capacities. Indeed, people's use of simulation is sensitive to these functional differences (for one example, see Dasgupta et al., 2018).

Before moving on, consider another way of assimilating learning through simulating to learning through observation: the position that simulation is a kind of learning through observation, but one that answers a different query and has a distinct target of observation. Whereas learning through observation answers queries *about the external world*, learning through simulation answers questions *about our internal models*. Correspondingly, the target of the former is the external world, while the target of the latter is ourselves.[7] On its own, this move is insufficient to accomplish our original goal: to explain how mental simulation can be used to learn about the world. In the water-pouring problem, the query answered through simulation was which glass would in fact spill first, not something about a mental model. But the view that simulation answers queries about our mental models could be supplemented to include an additional step: something like an inference connecting the internal observation to a proposition about the external world. For instance, in the water-pouring case, one would first observe that one's motor model "expects" the wider glass to spill first, and then infer based on the reliability of this model that the wider glass

[7] Magdalena Balcerak Jackson offers a version of a view on which simulation offers observations of the internal world, as opposed to the external world. On her account, the similarity in structure between imagining and perception explains why imaginings can give us phenomenal evidence (evidence about how things look). Observation (or in her terms, perception) has the role of providing phenomenal evidence as well as what she calls physical evidence (evidence about the way things are). Of these two roles, simulation can play the former, but not the latter. Other views similarly suggest that imagination can be a similar process to perception, but directed towards different subject matters. For example, Williamson (2016) takes the line that imagination gives us access to modal facts.

would in actuality spill first. In other words, learning through simulation could be reduced to a form of learning through observation (of one's internal model) followed by an instance of learning through inference (in which this internal observation figures in an inference about the external world).

We think this approach is problematic for several reasons. First, recall that we characterized learning through observation as a transition involving a relationship of external evidential support. This relationship would need to be modified to account for how one part of the mind could stand in the right (external?) relationship to another for this relationship (or a close analog) to obtain. The idea that one part of the mind performs a simulation, while another observes it, is uncomfortably homunculur, but perhaps not fatal. Our second worry involves the subsequent inferential step. An inference from an internal model to the external world would require (perhaps implicit) knowledge of what justifies the model's output. While some commitment concerning the relationship between the model and the world may often ground uses of simulation in science (see, e.g., Weisberg, 2012 on the role of models in science), it seems too demanding to require this of most mental simulation. As we elaborate below, it is characteristic of learning through simulation that we *lack* access to such commitments.

A final worry about this strategy is that it breaks up what looks like a single process into a series of fully distinct and epistemically independent steps. Doing so leaves us with the problem of explaining why these steps would be taken in sequence, and what relates them to each other. In sum, the strategy of reducing simulation to an internal observation followed by an inference seems to stretch our understanding of both observation and inference, and to miss what's epistemically distinctive about learning through simulation.

4. Simulation as Learning through Inference

An alternative way to approach simulation is to treat it as a form of learning through inference. By inference, we'll cast a wide net to include deductive and inductive reasoning, as well as abductive inference.[8] Crucially, inference here picks out a cognitive act rather than a relation between propositions, following Harman's (1986) distinction between inference and implication. As with the reduction to observation, this reduction would explain how simulation can result in learning by appealing to learning mechanisms that are better understood.

The project of reducing simulation to inference has several appealing features. First, like inference, simulation seems to rely on no external input; inference is the canonical form of learning from what you already know. A second promising element is that inference has a well-defined formal structure. Intuitively, what is distinctive about inference is that it involves the sort of process we can write down in the form of premises and a conclusion. If there were a parallel structure of premises and conclusion hidden in cases of simulation, this would be tremendously explanatory of the conditions for success and failure when it comes to learning through simulation.[9]

[8] On some views, abduction includes induction, and on others the reverse may be true.

[9] Norton (1991) argues along similar lines that thought experiments have a hidden argumentative structure, and so count as arguments for epistemic purposes. Gendler (1998) rejects the latter, instead holding that the supposed hidden argumentative structure does not capture the epistemic

Let's start with a definition of inference. While this is of course the subject of much dispute, most definitions share two linked components: (i) a requirement about the causal relationship between the thinker entertaining the premises and her entertaining the conclusion, and (ii) a requirement about awareness (or some weaker kind of reasons-responsiveness) that explains how the thinker could be described as endorsing the connection between premises and conclusions. Paul Boghossian's definition of inference illustrates both components:

> A transition from some beliefs to a conclusion counts as inference only if the thinker takes his conclusion to be supported by the presumed truth of those other beliefs.....S's inferring from p to q is for S to judge q because S takes the (presumed) truth of p to provide support for q. (Boghossian, 2014)

We're also interested in cases with more than one premise, and so we'll expand the definition to allow for a set of premises $p_1 \ldots p_n$ that jointly support the conclusion.

This notion of inference (as Boghossian himself notes) is too restrictive if we assume that the relevant sense of "taking the truth of *p* to provide support for *q*" requires *explicit endorsement* of a proposition about warrant. In an everyday context, the thinker usually tracks the support relation in a less explicit way. For instance, she might be said to implicitly infer on some basis when her use of inference is counterfactually responsive to that basis. That is, if her evidence had suggested that the basis did not hold, she would not have been disposed to make the inference (or, in other cases, if her evidence had implied that the basis holds, she would still have been disposed to make the inference).

Could a rat, on this implicit theory, count as inferring the location of a food source from an inductive generalization about past food locations? Since our aim is to pick out differences between simulation and inference across many kinds of thinkers, we'll adopt a permissive notion of inference that includes (non-human) animal thought. However, this definition will not capture *everything* that is sometimes called inference. Critically, we require the thinker to in some sense be in touch with the relationship between premises and conclusion, satisfying a weak form of condition (ii).

Suppose our rat's judgment that the food is on the left is caused by past experience of food being on the left when a particular odor is presented at the start of the maze. In a case of inference, he represents these experiences, and these representations cause his subsequent judgment (satisfying condition i). However, we require something more. While our rat can neither report the basis of his judgment nor his endorsement of some support relation between this basis and a conclusion, he can *act* as if the implicit form of condition (ii) holds: for instance, he might put two kinds of similar experience together to infer a generalization, such as "odors are only correlated with food position when they are presented at a particular time."

What would it mean to act as if he *wasn't* making a genuine inference? In this case, he might persist in turning to the left even when the context shifts, or fail to learn (or be responsive to) generalizations about his environment. Generalization and other forms of meta-learning are an

---

function of a thought experiment. See also Lombrozo (forthcoming) for further discussion of the limitations of reduction.

indication of inference because in order for this "learning to learn" (Harlow, 1949; Behrens et al. 2018) to occur, the rodent needs to do more than just be caused to go left based on some facts. What he needs is a sensitivity to the basing relation itself, which is to say he needs to infer.[10] This point is not merely theoretical; as one example, Tibbetts et al. (2019) present evidence that wasps make transitive inferences – i.e., they put past regularities together to learn a new one, which is exactly the kind of behavior that indicates true inference.

By defining inference in this way, we have a notion that's weak enough to capture implicit cases, but strong enough to identify why inference is a distinctive way of coming to know. Awareness of some particular propositions as the basis of one's belief (even if only in our very weak sense) explains why inference works: it's the kind of transition that builds on past knowledge in an accessible (and hence generalizable) way. We scrutinize inferential transitions, feel uneasy about those that proceed from a shaky basis, and evaluate how they fit together. Inference, unlike observation, makes salient a particular set of background beliefs (i.e., the basis) and their connection to the newly learned proposition.

Now, consider the difference between the original simulation of the two glasses in the water-pouring problem and the more explicit reasoning provided in Figure 1. The argument associated with the figure could be re-written as a set of numbered premises, and it seems natural that the person who uses this reasoning to discover the conclusion judges that the wide glass will spill first *because* she takes the truth of these premises to provide support for that conclusion.

On the other hand, something is missing in the simulation case. A thinker who uses simulation to answer the question is likely to find herself in the following position: she knows that her motor system in some sense "thinks" that the wide glass will spill first under the simulated conditions. However, she is unable to identify, recognize, or in any way point to what it is about these conditions, or indeed about her cognitive system, that generates the conclusion. She reached the conclusion because of *something* represented in her perceptual or motor system, and this causal relationship meets the first requirement (i) for inference. However, she cannot judge that the connection between premises and conclusion was one of support, because she is not in a position to identify the premises. She therefore fails to meet the second requirement (ii) for inference.

The issue here is not that the motor system is a holistic set of connected premises that support the conclusion - if so, this argument would also suggest that any kind of holistic inference is not real inference. Instead, the issue is that there are particular things about the inputs and cognitive processes behind this reasoning that support the conclusion, and *lots of things about the inputs and cognitive processes that have nothing to do with the conclusion*. But our simulator is unable to tell the difference, and thus to recognize or internally evaluate the relationship of support. To make this point even more obvious, consider that it might even be unclear to the simulator whether the answer came from her motor system, pure visual imagination, some form of spatial reasoning, or

[10] Siegel (2019) adopts an even weaker view of inference, on which it's the force of the conclusion that individuates inference from association or what she calls "mental jogging." We don't wish to rule out her theory as capturing the nature of inference as a cognitive kind. However, as we've defined it, inference is distinct from other kinds of learning, such as observation. On Siegel's view, perception itself is a form of inference, and so no such separation is possible. And so for the purposes of a taxonomy of coming-to-know, a more restrictive view of inference is required.

a combination of the above.[11] The critical point is that even if we assume that our thinker did not know whether her simulated output was coming from the visual system or required motor imagination, this does not undermine the fact that she learned the answer through simulation.

In fact, our simulator plausibly *knows* that she lacks the ability to identify the premises, and we can imagine that this bothers her. To rectify this, she might re-do the simulation, trying out what happens if the glasses are imagined to be very different in size, or if they are filled very high or very low. This re-simulation is a way to pull out the knowledge she lacks of what it is about the setup that matters to the output.

Our simulator's lack of knowledge might be seen as a flaw in simulation. And the inference-reductionist might use this to argue that simulation is instead a defective kind of inference. But this response misses something interesting about the process of re-simulation. The example above actually shows more than just a lack of knowledge in the case of simulation that is present in the case of inference. The further element is that re-simulation can help address this lack of knowledge. This feature implies that simulation and inference have different roles. After all, if inference requires identification of the premises, it cannot serve as a way to uncover those same premises - which is just what re-simulation does in our toy example. In Section 5 we expand on this thought to present our positive view.

5. What Simulation Lacks: Attribution and Warrant

In the previous sections, simulation did not prove to be easily assimilated to observation or to inference. Simulation of an event in some domain, unlike observation of a similar event in the same domain, typically has a role in calibrating and making explicit our own background models. Learning through simulation did not meet our definition of learning through inference because a typical simulation works without the thinker understanding which of her beliefs led to the output. Drawing on these features, we're now in a position to offer an account of what makes learning by simulation distinctive, and what this means for the epistemic function of simulation. In short, learning through simulation does not presuppose an answer to the following two questions:

> **Attribution Question**: Which inputs or features of the process led to the output?[12]
>
> **Warrant Question**: In virtue of what is the output justified?

While uncertainty about the answers to these questions is characteristic of mental simulation, it

[11] As it happens, motor simulation provides a more reliable basis for reaching the correct answer to the water-pouring problem than pure visual/spatial simulation, but both lead to better performance than soliciting a verbal response without a prompt to engage in some form of simulation (Schwartz & Black, 1999).

[12] In some cases, uncertainty about the answer to the attribution question might concern *how* the output was generated rather than what went into generating it. For instance, you might be told that your simulation used exactly these three pieces of information, but if you lack an understanding of how those facts could lead to the output, this would not resolve your real uncertainty. However, at least in some paradigmatic examples, "what" uncertainty is more diagnostic than "how" uncertainty.

may not be unique to simulation. However, we will argue that these forms of uncertainty partially differentiate learning through simulation from canonical cases of both observation and inference.

Our paradigm cases of learning through inference, and indeed the definition of inference, require that the reasoner have at least partial answers to these questions. She must know (or believe, on a fallibilist conception), that premises $P_1...P_N$ support her conclusion, and her knowledge of (or belief in) the conclusion must be based on this support. On a restrictive view of inference that requires explicit awareness of the relationship between the premises and the conclusion, inference presupposes attribution knowledge. And if this view requires explicit awareness of the inferential norms as governing norms, inference presupposes warrant knowledge.

However, recall that our account of inference was less demanding when it comes to explicit knowledge; On our view, inference can be defined functionally or behaviorally. Yet even on this weaker view, inference requires not only being moved by some particular premises, but having a standing disposition to follow rules connecting the premises and conclusion. The fact that this disposition is to *follow* the rules rather than merely accord with them allows us to attribute implicit warrant knowledge, even to our maze-running rat.[13] But we cannot do the same for our water-pouring simulator. By re-simulating the output, only sporadically relying on simulation, or even expressing confusion about how her simulation worked, the simulator behaves (or is disposed to behave) as if she does not have attribution or warrant knowledge. In contrast, by flexibly applying inference across the requisite situations, responding to defeating conditions on a particular inferential rule, and so on, the person making an inference behaves (or is disposed to behave) as if she did have attribution and warrant knowledge.

This "behaving as if" might seem like weak grounds to claim that she has or does not have knowledge of attribution and warrant, but the interlocutor who is pushing a thin, dispositional theory of inference should be more inclined to accept a correspondingly thin, dispositional ascription of knowledge.

Of course, a thinker who learns something from inference may understand very little about the conclusion. For instance, a student who completes a logical proof successfully using a conditional proof may not understand that her conclusion could have been reached by using a proof by negation.

---

[13] In the non-human animal case, a rodent who simulates a path through a maze during sleep may repeat that simulation over and over again with small variations before she tries out the path in real life. The rodent who makes a kind of inductive inference may also try out many different paths before coming to a determinate plan of action. However, our second rat will try out different paths systematically, in ways that represent her application of extracted general principles: for instance, she might carry out a kind of tree-based planning that displays an implicit awareness of the structure of anticipated events. This is in contrast to a kind of simulation oriented toward uncovering that structure, which could by definition not be so systematic. Of course, it's currently deeply controversial what repetitions of neural firings during sleep represent, if they represent anything at all. Our point here is that if replay during sleep is a kind of mental simulation, as Buzsaki et al. suggest, then we can distinguish the way the rodent treats these representations from the way the rodent would treat outputs of 'inferential' processes. And so even on the thin, dispositional notion of inference, simulation typically presupposes a further kind of uncertainty.

But unlike the case of simulation, she understands that her conclusion follows from her proof, and why - even if she can say nothing more specific than "because it logically follows from these premises."[14]

A similar move can explain why the answers to these questions are available in typical cases of learning through observation. Here, we know that we are making observations of the world, and this knowledge is critical for normal observation. On some intellectualized theories of observation, observers always endorse a proposition about the reliability of their own observations, which will straightforwardly imply attribution, since classifying a learning experience as observation is making an attribution of the content to the environment, and will straightforwardly imply warrant, since reliability is a kind of warrant.

But what about a less intellectualized theory of observation? Imagine that the correct epistemological view says that without any reflection, we have a strong default entitlement to rely on our perception, even absent any kind of demonstrable reason to do so (a position typically known as "perceptual dogmatism"). Even on this theory, we need to *go along* with our default entitlement to actually learn, so an observer who is uncertain about whether or not she is observing faces a problem. To learn through observation, she must go along with her perception and put aside the uncertainty. In asserting that the perceiver should go ahead and presume her perception comes from the external world even if she can't provide evidence that it's not actually an internally generated hallucination, she has attribution knowledge (or *acts* as though she does). In saying that the perceiver should presume her perception is reliable even if she can't give any evidence about the reliability of her perceptual faculties, she has warrant knowledge (or *acts* as though she does). That is, even on theories of observation that require no reflection or internal justification, observation requires attribution and warrant knowledge - it's just that this form of knowledge is ascribed very liberally.

6. What Simulation Offers: A Path to Self-Training

We've argued that uncertainty regarding attribution and warrant are part of what makes learning through simulation distinctive. But it may seem as though this feature is simply an epistemic weakness. After all, answers to the two questions may be valuable in many contexts, and even critical for being fully justified in relying on simulation.

On the contrary, we'll argue that this weakness can sometimes be a strength. Learning through simulation does not presuppose certainty about attribution and warrant, and this has two epistemically interesting effects: in the short-term simulation can allow learners to benefit from mental processes that are largely opaque, and in the long run simulation is a critical part of learning about attribution and warrant, thereby rendering those very processes less opaque. This explains why simulation plays dual roles: a one-shot simulation can provide a particular answer, and simulation in the long-run provides understanding of our own internal models which enables

[14] To exclude these cases, the Attribution and Warrant Questions must be understood as applying to the particular token output of the learning process, not to the learning process more generally, or to the type output of the learning process. These other kinds of interpretative uncertainty are present in many of our cases, but are not distinctive to simulation.

tuning and evaluation. We'll start with the long-run case.

Consider how, in the water-pouring case, a thinker might respond to her uncertainty about the scope and reliability of her motor model by re-simulating, as sketched earlier. To figure out how much a result depended on the initial setup, she might run the simulation again but this time imagining that the glasses were filled up very high or very low. This would help her learn that it doesn't matter whether the water is above or below the midline. She might re-check her motor simulations against a visualization, or even experiment on or observe the motion of liquids in containers in the real world. Or she could reason about what part of the setup would have led to the conclusion and what part would have been irrelevant. These responses are not alternatives to simulation, but next steps that build on her initial simulation. Through this process, she will begin to learn which parts of her motor and visual models lead to which outputs, when each model is likely to fit the world and when it ought not be relied on, and other related facts: that is, she will resolve attribution and warrant uncertainty.

This process of uncovering the workings of the model is central to prominent empirical theories of mental simulation. For instance, in the motor case, research finds that expertise with one sport can result in better pattern recognition in *another* sport. This isn't because individual motor skills transfer from one domain to another (in general such transfer is weak), but rather because expertise involves a kind of meta-cognitive knowledge about the appropriate uses of simulation – the experts have learned something about attribution and warrant when it comes to the use of motor imagery for skill improvement.[15]

In such cases, expertise in motor skill improves the outputs of internal models, but it also renders an aspect of the model available for evaluation, which in turn provides new opportunities for training. Training requires holding up a part of the model for evaluation, and greater accessibility means that evaluation of the model and its output can draw on a wider body of evidence. Evaluation, of course, subsequently allows the model to be tuned and adjusted, and for its appropriate scope to be understood. And so over time, this iterative process resolves warrant and attribution uncertainty as the internal model is built and trained. To simplify somewhat, this is because attribution uncertainty is an accessibility problem, and warrant uncertainty is an evaluation

[15] Comparing motor transfer with cognitive transfer, Schmidt and Young (1987) summarize the state of the research: "When such measures are applied to experiments on motor transfer, the outcomes arc relatively consistent Motor transfer is generally very small" (pp59). So one might expect that motor experts in one domain would not have any significant advantage in another related but distinct domain – and might even perform worse due to interference. Abernathy et al. (2005) found that to the contrary, experts in different sports performed significantly better than novices at pattern recognition and classification in a different sport from that of their expertise. Williams et al. (2006) compared expert and non-expert soccer players, finding that experts relied more on structural features where non-experts relied on superficial features. MacIntyre et al. (2014) theorize that these differences stem at least in part from a link between mental practice and metacognition: experts have enhanced abilities to imagine and simulate, and this is connected to a greater awareness of the appropriate use of simulation: "experts may simply possess greater meta-cognitive knowledge of how to employ imagery effectively for skill improvement as compared to novices."

problem – and as we've seen, the two problems are tied together such that better accessibility makes better evaluation possible.

We'll now argue that unlike learning through simulation, learning through observation and learning through inference do not offer systematic ways of resolving warrant and attribution uncertainty. By "systematic" we intend to allow that sometimes, this uncertainty can be resolved by luck, or by brute external alteration such as a kind of re-programming, or even occasionally by a stroke of good reasoning or observation. But we hold that overall, neither observation nor inference are well-positioned for this role as compared with simulation. This argument will therefore presuppose that the thinker already has a need for simulation because of warrant and attribution uncertainty; we'll take up the question of why a thinker would ever end up with this kind of uncertainty in Section 8.

Consider a case discussed by Tal (2011). On his analysis, the use of simulation allowed physicists to confirm the Bose-Hubbard model in the following way. First, a series of observations were made concerning phase transitions, including one that seemed to fit with the Bose-Hubbard model (in Munich in 2001), and one that seemed not to fit with the model (in Zurich, 2004). After this, a third group of physicists ran a computer simulation that produced simulated predictions for what would be observed if the Bose-Hubbard model were true. The result of this simulation revealed that the second set of experiments from Zurich, which initially seemed to be at odds with the model, were actually the sort of result that we would expect to see on the Bose-Hubbard model. For our purposes, the important feature of this case is that the physicists already had the observations that would confirm their theory when they ran the simulation – but it was only after analyzing the simulation that they understood these observations as evidence in favor of the theory. Therefore, the epistemic function of the simulation could not have been filled by observation, since the observation had already occurred.

Now, a perfect physicist might use impeccable inference to determine analytically the consequences of her theory. But the physicists studying the Bose-Hubbard model were presumably unable to do so; this is precisely the sense in which they were uncertain prior to the simulation. So just as with observation, inference cannot be a systematic solution to this problem. This is because physicists who could infer their way reliably from the theory to its various predicted observations would never be in this situation in the first place. Therefore, neither of the other learning methods will in general resolve warrant and attribution uncertainty.

The preceding argument concerned the long-run use of simulation. In the one-shot case, our account suggests that thinkers can learn from the output of a single simulation, but this form of learning is fragile given the underlying uncertainty. One-shot reliability will be imperfect since warrant uncertainty prevents the thinker from being perfectly responsive to the scope and meaning of the output, unless by luck. In the one-shot water-pouring case, for example, when you began the simulation, you were either already justified in believing the output of the motor system or not. However, you did not yet believe the wide glass would spill first before carrying out the simulation. It would even be controversial to say that you implicitly believed it, since you would be unlikely to consistently behave as though it were the case, and your motor system may not even have represented this fact about the two glasses so much as other features of liquids that imply the wider

glass will spill first.[16] And so the simulation led you to believe the wide glass would spill first, and thereby to know it. Thus representational change associated with simulation enables reliance on one-shot simulation, but does not provide a new source of justification so much as enable access to a potential source to which the thinker is in some sense already connected. In contrast, in the long-run case, representational change to the model and its accessibility is an integral part of the thinker coming to have genuinely novel justification for her subsequent beliefs.

Stochasticity in simulation also functions differently in the one-shot case than in the long-run case. While not all simulations are stochastic, most scientific simulations seem to be, and it is standard to model mental simulation as a kind of non-deterministic sampling (see Zhang et al. 2012, and Bramley et al., 2018, for two examples in different domains). Stochasticity looks like a liability in one-shot simulation, since the output may be an unlikely one that mischaracterizes the internal model and the target system. Rather, it's only over the long-term, or with repeated one-shot simulations, that stochasticity emerges as a desirable feature of simulations, allowing the system to encode uncertainty while maintaining specificity.

The combination of these two roles for simulation based on time-scale suggests an interesting consequence: because the long-run role changes the properties of the model over time, the one-shot uses of simulation might depend on where they occur in the long-run process. In general, one-shot learning will be more reliable later on in the process.

We're now in a position to see another benefit of long-run simulation related not to the process itself, or to its inputs or outputs, but to what triggers the use of simulation in the first place. Compare an expert simulator with a novice. The expert will be triggered to employ simulation more appropriately, and likely more often, than the novice; because her model makes more explicit and accurate predictions, the impetus to simulate will arise in contexts where her predictions are relevant, whereas the novice will have fewer and less explicit predictions that will relate to different circumstances precisely because they are less well calibrated. Of course, many uses of simulation may still be completely initiated by an external factor, such as a teacher asking you to solve a puzzle by imagining the pieces, or the authors of a paper asking you to consider the water-pouring problem. But in many critical domains, successful simulation depends on simulating under the right conditions. And as the simulator becomes more and more knowledgeable, she will get closer and closer to meeting our definition of inference, at least in its weaker dispositional form.

7. How Simulation Works: Changing accessibility conditions through representational extraction

In Section 6, we suggested that simulations make information more accessible. In the long-run case, changes in accessibility enable enhanced evaluation and learning over time, whereas in the one-shot case, accessibility was not epistemically significant in terms of justification, but

[16] Of course, we can easily imagine a case where you *do* implicitly believe the output of the simulation even before you simulate, and then later come to explicitly believe it. In this case, you would not come to know by simulation since you already knew, but we might say that you now believe more firmly or understand better. Likewise, we've allowed for simulations that have only implicit representations as their output. How to treat these cases would depend on the theory of implicit knowledge.

nonetheless played a role in explaining how you came to have knowledge of *q*. It might therefore be tempting to characterize learning through simulation as a process by which implicit knowledge becomes explicit knowledge, and center the epistemic function of simulation on this transition. We will suggest that this characterization is not quite right.

The implicit/explicit characterization is compelling because it explains how knowledge is possible at all (it was there all along), and also what changes through learning (something formerly implicit becomes explicit). For instance, in the water-pouring problem, the answer was *in some sense* already encoded in the learner's perceptual or motor system. On the other hand, one result of the simulation is that the answer is now encoded in a form that can be expressed in words, used as a premise in further reasoning, and so on.

The idea that a mental simulation can render implicit knowledge explicit is reflected in claims about thought experimentation in both philosophy and psychology. For example, Mach (1897, 1905) suggested that thought experiments reveal "instinctive knowledge." Clement (2009) writes that mental simulations can "draw out implicit knowledge" that can then be described in linguistic form. We think this idea gets something right, but also faces serious limitations.

First, outside a handful of specialized literatures within psychology, it's unclear what it means for something to be implicit or explicit. Are these claims about the structure of mental representations? About the mechanisms by which they are accessed? And what is it that changes through learning? Ideally, an account of learning through simulation should support answers to such questions.

Second, some cases of learning through simulation seem to involve transitions *within* the realm of the implicit, or *within* the realm of the explicit. For instance, in the rodent example provided by Buzsaki, an implicit representation about possible head directions could be made available to an implicit module encoding memory for past navigational trajectories. Within the realm of the explicit, the physics modeler's computer simulation could result in a transfer of explicit general assumptions about bosons in a lattice into explicit predictions about what particular signals will be detected by a scanning device. A better way of describing this change in availability is in terms of a change in the *accessibility conditions* for some information, where this change occurs via *representational extraction* through simulation. Below we unpack these ideas.

Following Lombrozo (forthcoming), we make the relatively uncontroversial assumption that "different mental representations are available to different mental processes." With this assumption in place, we can see that the output of a mental simulation will be available to some processes, but not to others. For example, the output of a particular motor simulation might be available to guide an arm movement, but not to verbally report; the output of a visual simulation of the water-pouring problem, on the other hand, may well be available for verbal report, and also as a premise in further reasoning. The crucial observation is that a mental simulation will often generate a representation – the output – that makes information available to a system in a new way. Prior to the mental simulation, a simulator's motor system may have encoded a regularity between glass width and pouring angle, but not in a form that was available for verbal report or further verbal reasoning.

This approach goes beyond an implicit / explicit distinction in recognizing a much broader range of *accessibility conditions* for information.[17] We can posit a range of mental processes, each of which imposes some constraints on what it will accept as "input." A mental simulation, like inference, produces "new information" in only a limited sense; instead, what it offers is a change in accessibility conditions, and hence in how existing information can be deployed. The result is what we call "representational extraction," which can occur in both one-shot and long-run cases.

In a case of one-shot simulation, information that may be encoded only implicitly in the process by which a simulation operates can constrain the output of the simulation. As a result, downstream systems will benefit from this information indirectly – by virtue of access to the output, which has been shaped by the relevant information.

Turning to the case of long-run simulation, we can see an additional way in which simulation can "extract" a new representation. As simulation is repeated and the simulator begins to resolve her uncertainty concerning attribution and warrant, she can come to represent the dependencies, relationships, and other content of the internal models that underwrote the simulations. Most often, this process will also involve inference. For example, she might infer a particular dependence relationship from repeated pairs of initial conditions and outputs, just as we described the water-pourer as playing around with water levels to see how changes would affect the output. Even if the internal workings of the model started out fully opaque, this backwards inference would increasingly ground her knowledge of the relationships encoded in the model.

Representational extraction explains the puzzle posed by the Hallucinated Replay example: how could less accurate simulated "observations" be more useful than more accurate genuine observations? The answer is that simulation, unlike observation, results in representational extraction – the simulated "observations" extract, or make available, consequences of the model that were previously only represented in a more limited way. Without extraction, self-training is necessarily limited – as we've argued, it is only because parts of the model are exposed to evidence that they can be improved. In Hallucinated Replay, the agent could "improve" parts of the model that could not ever be trained by experience, since they predicted what would happen after an impossible state of the world. But extracting even these steps has a use, since these impossible transitions are related to other parts of the model and partly determine higher-order generalizations. Representational extraction is by its nature a way of making a piece of one's representation exposed to a wider range of one's evidence, a process that has not merely intellectual value but

[17] Philosophers are increasingly recognizing that accessibility of information is a major determining factor in producing actions, guiding reasoning and every other cognitive activity: it's not just what you believe that matters, but which beliefs you are disposed to access in a given context. Harman (1986) drew attention to relevance and access as central to changes in belief (see also Stalnaker (1991)) and more recently Elga and Rayo (ms.) provide a semantic framework for representing accessibility relationships. These theories essentially take on the question of how limitations in accessibility should be understood, and how it is rational to respond in light of these limitations. Our approach is consistent with these ideas, but considers a slightly different question: we are concerned not (only) with the conditions under which a belief is accessible for explicit report or reasoning, but more broadly in the conditions under which a *representation* (which need not be a belief) is accessible to a *mental system* (which need not be verbal report or reasoning).

practical pay-off.

8. What Kinds of Learners Need to Simulate?

We have argued that observation and inference don't make good substitutes for simulation. In doing so, we have described thinkers who have uncertainty about their own internal models that only simulation can address. However, this uncertainty need not characterize every thinker. So we are now in a position to ask: what kind of creature would need to simulate in the first place? Or, what kind of learning environment would make simulation necessary?

Throughout this paper, we've contrasted simulation to inference and observation. This contrast reveals two conditions that must be met for simulation to be necessary (in the sense that it cannot be replaced by observation or inference):

1. The thinker must be limited in the kind of information she has free access to. That is, she sometimes cannot perform a particular experiment or make a particular observation without paying a significant cost or incurring a significant risk.
2. The thinker must have some representational opacity.

Condition 1 makes it so that she cannot substitute experimentation for simulation, and Condition 2 makes it so that she cannot substitute inference for simulation. And our analysis of the distinctive uncertainty that rationalizes simulation allows us to see these conditions as following directly from the function of simulation. If the agent could experiment as much as she desires, contra Condition 1, she could figure out exactly how reliable her various internal processes are. That is, she could resolve warrant uncertainty. And if the agent were totally transparent to herself, contra Condition 2, she would be able to resolve any attribution uncertainty. Together, the two conditions make simulation a vital capacity.

All of us live in environments with limited evidence, and thereby meet Condition 1. Even the most idealized agents are presumably still subject to this limitation – and in fact, even with abundant evidence, limitations in cognitive capacity might create trade-offs between collecting evidence and making plans. Momennejad et al. (2018) think of offline simulation as crucial for this reason: it allows the agent to save time and cognitive resources at the moment of decision.

But why would a thinker have representations that are opaque to her? Or in other words, could there be a reason that some mental processes are at least initially a black box? One kind of opacity springs from modularity, even in weak forms. If the thinker is made up of specialized modules that each employ their own representational format, it might be costly to translate between them. Further, there might be no lossless means of translation. In either case, this would result in some content in one module being inaccessible to another module or cognitive process, as we have already suggested.[18]

[18] Opacity could arise even in agents who are only very weakly modular – for instance, a thinker who shifts between two styles of thought (depending on context) that broadly recruit the same resources. This thinker may not be able to access both ways of thinking in every context, and yet the capacity to simulate might allow her to jump between contexts by setting up a virtual version of another context.

At this point, we can return to the question of the relationship between simulation in science and mental simulation.[19] Our account appeals to epistemic features (warrant and attribution uncertainty) that can be applied to a thinker, a scientific community, or even to other kinds of group agents. Throughout the discussion, we've drawn attention to common features between simulation in these disparate contexts, such as the propensity to re-simulate. Although representational opacity might emerge for different reasons at individual and group levels,[20] in both cases simulation can reduce opacity through representational extraction that changes accessibility conditions: information becomes available to other parts of the mind, or to the human operators of machine simulations. We therefore expect the core features of our account to apply to learning through simulation quite broadly.

9. Conclusion

This paper has addressed the question: how can we learn through simulation? After considering, and rejecting, models of learning through simulation that treat simulation as a kind of observation or a kind of inference, we have presented a theory on which simulation is distinct from either of these better-understood forms of learning. Simulation does not presuppose an understanding of how the output was generated (attribution) or how the process of generation should be relied on (warrant). However, simulation over time makes progress on resolving this uncertainty through representational extraction.

Our account sheds light on our starting point: the idea that simulation is somewhat like observation and somewhat like inference. Simulation is somewhat like observation in that the thinker gets the output "from outside." This is because simulation involves representational extraction, or the bringing in of information from one cognitive system or model to another. Simulation is somewhat like inference in extracting something from what we already have. Moreover, as we simulate more and more, and get more and more evidence from other sources, our internal model goes from a work-in-progress to, hypothetically, a complete product. Simulation aims at resolving attribution and warrant uncertainty, and in the limit when the uncertainty is fully resolved, simulation becomes a fully transparent way to process information through a fully trained internal model. In other words, simulation becomes inference.

---

[19] See Bratman (2013) or Gilbert (2000) for a related discussion.

[20] Representational opacity will potentially be instantiated differently across individual and group agents. For instance, in the Bose-Hubbard case, representational opacity might result from the fact that the computer model contained latent information about the likelihood of various observations based on the theory – and this information was opaque to the scientists debating the various theories. But we might also consider the information to be represented in the Bose-Hubbard theory itself, rather than the computer model. Addressing this question, and other related issues, would require a theory of the group thinker, or the group of thinkers, that is beyond the scope of this paper.

Acknowledgements.

We would like to thank Adam Elga, Thomas Icard, Josh Knobe, Eric Mandelbaum, Shaun Nichols, and Shannon Spaulding for invaluable comments. Thanks also to members of the Concepts and Cognition Lab and attendees of Princeton's Parallel Distributed Processing workshop.

References.

Atkeson, C. G., & Santamaria, J. C. (1997). A comparison of direct and model-based reinforcement learning. In *IEEE International Conference on Robotics and Automation*, (Vol. 4, pp. 3557-3564). IEEE.

Boghossian, P. (2014). What is inference?. *Philosophical Studies*, *169*(1), 1-18.

Bramley, N. R., Gerstenberg, T., Tenenbaum, J. B., & Gureckis, T. M. (2018). Intuitive experimentation in the physical world. *Cognitive psychology*, *105*, 9-38.

Bratman, M. E. (2013). *Shared agency: A planning theory of acting together*. Oxford University Press.

Brown, James Robert and Fehige, Yiftach, "Thought Experiments", *The Stanford Encyclopedia of Philosophy* (Summer 2017 Edition), Edward N. Zalta (ed.),

Buzsáki, G., Peyrache, A., & Kubie, J. (2014). Emergence of cognition from action. In *Cold Spring Harbor Symposia on Quantitative Biology* (Vol. 79, pp. 41-50). Cold Spring Harbor Laboratory Press.

Dasgupta, I., Smith, K. A., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2018). Learning to act by integrating mental simulations and physical experiments. *bioRxiv*, 321497.

Elga, A., & Rayo, A. (2016). Fragmentation and information access. *Draft manuscript, MIT/Princeton*.

Gendler, T. S. (1998). Galileo and the indispensability of scientific thought experiment. *The British Journal for the Philosophy of Science*, *49*(3), 397-424.

Gilbert, M. (2000). *Sociality and Responsibility: New Essays in Plural Subject Theory*. Rowman & Littlefield Publishers.

Hamrick, J. B. (2019). Analogues of mental simulation and imagination in deep learning. *Current Opinion in Behavioral Sciences*, *29*, 8-16.

Harman, G. (1986). *Change in view: Principles of reasoning*. The MIT Press.

Jackson, M. B. (2018). Justification by Imagination. In F. Dorsch / F. Macpherson (Ed.) *Perceptual*

Imagination and Perceptual Memory. Oxford University Press.

Jeannerod, M. (2001). Neural simulation of action: a unifying mechanism for motor cognition. *Neuroimage*, *14*(1), S103-S109.

Lombrozo, T. (2017). "Learning by thinking" in science and in everyday life. In P. Godfrey-Smith & A. Levy (Ed.), The Scientific Imagination . Oxford University Press.

Mach, E. (1883). *Die Mechanik in ihrer Entwicklung: Historisch-Kritisch Dargestellt*, seventh edition.

—— (1897). "Über Gedankenexperimente", in: *Zeitschrift für den physikalischen und chemischen Unterricht*, 10: 1–5.

Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2018). Offline replay supports planning in human reinforcement learning. *eLife*, *7*, e32548.

Norton, J. (1991). Thought experiments in Einstein's work. *Horowitz and Massey, 1991*, 129-148.

Schwartz, D. L., & Black, T. (1999). Inferences through imagined actions: Knowing by simulated doing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(1), 116.

Siegel, Susanna (2019). Inference Without Reckoning. In Brendan Balcerak Jackson & Magdalena Balcerak Jackson (eds.), *Reasoning: New Essays on Theoretical and Practical Thinking*. Oxford University Press. pp. 15-31.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Dieleman, S. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484.

Tal, E. (2011). From data to phenomena and back again: computer-simulated signatures. *Synthese*, *182*(1), 117-129.

Weisberg, M. (2012). *Simulation and similarity: Using models to understand the world*. Oxford University Press.

Wellman, H. M. (1992). *The child's theory of mind*. The MIT Press.

Zhang, J. , Hedden, T. and Chia, A. (2012), Perspective-Taking and Depth of Theory-of-Mind Reasoning in Sequential-Move Games. Cognitive Science, 36: 560-573. doi:10.1111/j.1551-6709.2012.01238.x