e-ISSN 2723-9640

JoLMA
Vol. 6 — Num. 1 — June 2025

Rethinking the Simulation Theory

Robert M. Gordon
University of Missouri, St Louis, USA

Abstract This paper revisits the Simulation Theory (ST) as a framework for understanding human social cognition, challenging traditional 'theory of mind' or 'folk psychology' approaches. While these theory-based models posit that humans use an implicit body of knowledge to interpret and predict others' behavior, ST emphasizes the use of mental simulation, leveraging the brain's existing mechanisms for planning and prediction. By employing a predictive coding strategy, the brain minimizes cognitive load, interpreting others' actions through 'inverse planning' – a process that reuses one's own action planning system to hypothesize the goals and intentions of others. The concept of agent-neutral coding is introduced, proposing that inputs for self and others are initially shared, reducing the need for explicit mental state attributions. This approach not only economizes cognitive resources but aligns with evolutionary perspectives on human social interaction in small, cohesive groups. In addition, the paper explores the role of perspective-taking and error correction in adapting shared mental representation. This reevaluation of ST underscores its efficiency and adaptability, offering a streamlined alternative to theory-based accounts of social cognition.

Keywords Simulation theory. Theory theory. Social cognition. Other minds. Primate evolution.

Summary 1 Introduction. – 2 The Thirst for Efficiency: The Predictive Approach. – 3 Inverse Planning. – 3.1 Simulative Inverse Planning. – 4 Agent-neutral Coding. – 5 Perspective-taking and Positional Correction. – 6 An Evolutionary Perspective. – 7 Ignorance And False Belief. – 8 Knowledge First. – 9 Emotions. – 10 Conclusion.

Edizioni Ca'Foscari

Peer review

Submitted 2025-01-18
Accepted 2025-08-01
Published 2025-09-01

Open access

Citation Gordon, Robert M. (2025). "Rethinking the Simulation Theory". *JoLMA*, 6(1), 51-68.

DOI 10.30687/Jolma/2723-9640/2025/01/003

1 Introduction

Since the 1960's it was widely assumed that human competence in interpreting and anticipating the behavior of others depends on a body of general knowledge – a theory, commonly called 'folk psychology' by philosophers and 'theory of mind' by psychologists. This is a theory, we can say colloquially, of 'what makes us tick.' And the core stipulation of the theory is that behavior, or at least intentional action, is caused by the agents' beliefs and desires primarily, although other mental states may be added to the mix. Our capacity to theorize about these underlying states is sometimes called mentalizing; but the term is often interpreted more broadly, to cover our everyday understanding of others' behavior, with or without reference to mental states.

In the 1980s, the 'simulation' theory (ST) posed the first serious rival to the 'theory theory' (TT). ST locates the main source of mentalizing competence in a procedure or set of procedures called 'simulation,' or 'mental simulation.' Introduced by philosophers (Gordon 1986; Heal 1986; Goldman 1989), this account is usually thought to challenge the very assumption that mentalizing is an application of an implicit theory of mental states. The TT- ST debate soon became a topic of interest to developmental psychologists and others working on social cognition.

One of the initial motivations for a simulation account was efficiency, or getting the most benefit with the least expenditure of resources – in brain circuitry, processing time, and metabolic energy. The importance of getting the most with the least, as well as the computational means for accomplishing that, have become clearer in the past two decades.

What would simulation offer? For one thing, simulation would spare the brain the overhead costs of acquiring, storing, and applying the information needed to construct a model of what makes us tick. But an important part of the simulationist response was to ask a simple question: Why would a system need to invest in a general theory or model of systems like itself? Wouldn't it be more economical simply to use itself as a stand-in for these other, similar systems? This question will be a major theme in what follows.

2 The Thirst for Efficiency – The Predictive Approach

How is it possible for the human brain to make sense of the complexities of human behavior and interaction? Consider everything you know about the behavior and interactions of everybody you know or remember. How can all this be handled by the primate brain – in fact, just a portion of the primate brain?

Part of the solution would be to make that portion, namely, the neocortex, much bigger. This is clearly what happened in the evolution of the primate brain. According to Robin Dunbar's widely accepted Social Brain Hypothesis, primate societies are unusually complex, and the need to manage such complexity is the main explanation for the fact that primates have unusually large brains. Primate sociality is based on bonded relationships that underpin coalitions, which in turn are designed to buffer individuals against the social stresses of living in large, stable groups. This is reflected in a correlation between social group size and neocortex size in primates (but not other species of animals). The correlation, in humans at least, is due to our *mentalizing skills* as they grow to encompass an expanding community.

In the light of this thirst for resources, it is disappointing to note that many proponents, as well as most critics of the simulation theory, have supposed simulation to be an elaborate set of processes involving recognition of one's own mental states and an implicit inference from oneself to others. The form of inference is essentially the old argument from analogy, which requires that one first introspect in order to recognize one's own mental states under various conditions; then, after identifying those states, inferring that the other is in similar states.

Why the need for this elaborate, intellectually loaded process? Why all the judging and recognizing and leaps of analogy? Is there something about the primate brain that demands all this thinking? I don't think so.

The approach presented here is much in line with the current view in psychology and neuroscience that neural systems tend to reduce metabolic and other expenses by employing a predictive coding strategy. This is a strategy of 'guessing ahead.' Rather than waiting for the world to bombard us with new information, the system makes its latest best guess as to what will be coming in. This process of predicting input values minimizes the need for new information input, in that only discrepancies, or information that conflicts with the predicted values (prediction errors), need be encoded.

Any corrections or departures from this default are likely to be relatively small, requiring minimal resources to encode these differences.

Thus understood, simulation would resemble schemes commonly used in the digital transmission and storage of video content (Gordon 1992). Typically, little or no visual content changes in, say, the thirtieth of a second that separates one frame from the next; successive frames in a video sequence are nearly always very similar. Therefore, it is an efficient strategy to treat each frame initially as a copy of the previous frame – and then looking for any discrepancies, a much smaller task than building an entire new frame from scratch.

A comparable simulation account will show how our mentalizing system exploits massive redundancies to achieve extreme code compression. Rather than building from scratch a picture of the other's reasons and motives, we start with – not ourselves, strictly speaking, but the world.

3 Inverse Planning

To address the question of how the brain interprets the observed actions of others, Baker, Tenenbaum, and Saxe (2006) suggest that we adopt a framework that has been particularly fruitful in studies of vision. Contrary to the widely held view that visual perception simply pastes together a complex scene from elements such as lines and edges, it is now thought that the process works in reverse. Our brains generate predictions about the incoming sensory information based on past experiences and learned patterns. These predictions help anticipate what we are likely to perceive in a given situation. When actual sensory input matches these predictions, the brain processes the information more efficiently, leading to a sense of familiarity and reduced cognitive load. However, when there are discrepancies between predictions and actual input, our brains update their models to better match the current environment. This theory highlights the active role of the brain in shaping our perception of the world around us. The interpretation of a visual scene might involve, essentially, using in reverse the process of *producing* such a scene. Analogously, the interpretation of another's behavior might be understood as a comparable inverse problem (Baker, Saxe, Tenenbaum 2011; Baker, Saxe, Tenenbaum 2009):

> By analogy, our analysis of intentional reasoning might be called 'inverse planning', where the observer infers an agent's intentions, given observations of the agent's behavior, by inverting a model of how intentions cause behavior. (Baker, Tenembaum, Saxe 2006, 100)

The process is *inverted* in that, instead of proceeding forward from a given intention to its behavioral execution, it takes the behavior as the given and determines the intention most likely to have produced it. The planning process would thus be used as a mechanism for testing hypotheses about underlying intentions.[1]

1 In the broadest terms, inverse planning exemplifies hypothesis-testing as unconscious inference, an idea introduced in the perceptual realm by Helmholtz (1856). The proposal bears some resemblance to 'hypothetico-practical' inference (Gordon 1986),

Strictly speaking, however, the term 'inverse planning' suggests that the very mechanism that is used to plan our own behavior may be reused as a platform for testing hypothetical explanations of the observed behavior of other agents. This would, in effect, be a way of *simulating* ways of generating the behavior. However, Baker, Saxe, and Tenenbaum (2011) actually propose something more complicated. The authors speak of inverting a *model* or *theory* of the planning process. As they point out, their project originated as an attempt to formalize an intuitive theory of mind thought to underlie our interpretations of behavior – the so-called 'theory theory':

> On a theory-based interpretation, inverse planning consists of inverting a causal theory of rational action to arrive at a set of goals that could have generated the observed behavior. (Baker, Saxe, Tenembaum 2009, 347)

The theory-based approach attributes to the brain a capacity for detachment: it *stands back from its own operations* and employs instead a general theory or model of these operations. As distinct from actual action-planning, the theory theorist proposal is that in mentalizing about others the brain engages in *plan-theorizing*, *theorizing about* the steps in the other's planning process. The proposal assumes that we humans have an intuitive theory of mind and that our brains employ this theory not only in our explicit attributions of mental states but also in their unconscious subpersonal neural processing. I will call this *inverse plan-theorizing*. Thus understood, it does not make use of our capacity for planning: it is not inverse planning as such, i.e., an inverse reuse of one's own action planning system. Strictly speaking, what would be analogous to 'inverse graphics,' where perception involves searching among alternative hypothetical ways of building a scene, would be 'inverse planning,' understand as a search among alternative hypothetical ways of generating (planning) an action to find the most plausible simulation of the planning that might have generated the observed action.

Baker, Saxe, and Tenembaum acknowledge that a simulation-based account would cover the data just as well as their theory-based account:

> On a simulation account, goal inference is performed by inverting one's own planning process – the planning mechanism used

modeled on a traditional model of the scientific method, hypothetico-deductive inference. Instead of forming hypotheses and *deducing* consequences that match observations, hypothetico-practical inference would form hypotheses and then *act on* them, *producing* consequences that match the observed behavior of the other agent.

> in model-based reinforcement learning - to infer the goals most likely to have generated another agent's observed behavior. (Baker, Saxe, Tenembaum 2009, 347)

If indeed such reuse of its own 'first-person' planning system would be sufficient for goal inference, the question arises: Why would the brain need to operate instead on a model of the planning process? Here again, using an existing system would avoid the overhead costs of storing and utilizing an information-rich theory or model. Moreover, first person inverse planning would seem to be the proper analogue of the inverse graphics account of vision. As inverse graphics is the inversion of a causal physical process of scene formation (Baker, Saxe, Tenembaum 2011), so inverse planning should be the inversion of *a physical process* of action determination – *not* the inversion of a causal *theory of* a physical process of action determination. The 'vision is inverse graphics' idea is generally understood to be an analysis-by-synthesis paradigm, and analysis by synthesis is not analysis by a *theory of* synthesis.

3.1 Simulative Inverse Planning

There is at least one crucial difference between the simulation account of inverse planning (where the planning process itself is inverted) and the theory-based account (where a model of that process is inverted). On the simulation account, one and the same action planning system has a double function: in addition to its primary use in generating one's own actions, a reuse, or secondary use, in which the planning process is inverted in order to infer the goals and reasons that lie behind another agent's observed behavior. Moreover, it appears likely that the secondary use of the action planning system, namely, inverse reuse for explanatory purposes, runs concurrently with its primary use, for generating one's own actions. Otherwise, we would have to suspend our own actions in order to interpret the actions of others. Thus, the system is translating existing inputs into action and at the same time looking for hypothetical inputs that would explain the perceived actions of others. Concurrent processing for self-action and other-understanding would be consistent with evidence of 'motor contagion,' or interference effects between observed and executed actions. First noted in the case of biological movements, it has been suggested that motor contagion may be "the first step in a more sophisticated predictive system that allows us to infer goals from the observation of actions" (Blakemore, Frith 2005, 260). Indeed, recent research indicates that such interference is markedly increased when the observed movement is directed toward a visible goal (Bouquet et al. 2011). This interference suggests a competition for resources,

and thus that the same, or strongly overlapping, neural resources are employed concurrently in goal-directed action planning and in interpreting the goal-directed actions of others.

Such concurrent double employment raises the question: What, if anything, must *change* as the planning system switches from primary use to reuse, and from self to other? Specifically, what happens to the existing inputs? When the system switches to inverse planning as it seeks to explain another's behavior, does it clear the slate and approach the task with no a priori top-down commitments? More specifically, for the inverse use, does the brain suspend the beliefs, desires, preferences, emotional valences, affordances, and other influences on one's own action planning? That is, does it expend energy to intervene and wipe away the inputs and start with a blank slate when simulating others? That would seem wasteful both in loss of information and in use of resources. At the opposite extreme, does the brain leave all inputs in place, add no others, and seek the best explanation of the other agent's behavior strictly on the basis of the beliefs, desires, preferences, emotional valences, affordances, and other influences on one's own action planning? That would seem highly limiting. The most plausible account would be for the brain to default to this do-nothing position and devote its limited energy to looking for problems. Focusing on exploiting redundancy and then checking for exceptions is much in line with a widely held view in cognitive science: that neural systems tend to reduce metabolic and other expenses with a predictive coding strategy (Clark 2013). As in the case of vision, this is a strategy of 'guessing ahead.' Rather than waiting for the world to bombard us with new information, the system makes its latest best guess as to what will be coming in. This process of predicting input values minimizes the need for new information input, in that only discrepancies, or information that conflicts with the predicted values (prediction errors), need be encoded.

4 Agent-Neutral Coding

Gordon (2021) argues that the top-down inputs to inverse planning would default to *agent-neutral coding*. That is, inputs, including factual inputs, would by default remain the same for self and other; that is, the same unless corrected, e.g., in response to predictive error. Coding begins as agent-neutral, in the sense that any differentiation would be the result of intervention of some sort: Identical coding for self and other would be the default. With agent-neutral coding, one's own actions and the actions of others are constrained by the same inputs unless there is reason for differentiation. The claim is not that *my* inputs are carried over, but rather that an *undifferentiated* input, neither mine nor the other's, becomes differentiated into mine and

the other's. It is of course my own mental states that provide input to the forward planning of my own actions, and it is representations of the other's mental states that feed into the inverse use of the planning system to explain the other's behavior. It might be supposed that the system has to distinguish these in some way. But this is not so. Unlike intentions and motor plans, beliefs may remain happily undifferentiated, and failure to differentiate is not only not pathological, it is the norm. What the system needs to 'know' is, simply, that there is a puddle in the path; it can deal with undifferentiated, impersonal 'facts,' without marking them as facts-to-me, facts-to-you, or facts-to-another – or, in other words, as facts *as I believe them to be*, or you, or another. Moreover, as will be argued, simple 'factive' explanations, such as, 'She stepped to the side because there was a puddle in the path', are the preferred form of action explanation, in contrast to 'because she believed…' explanations (use of 'because she believed…' is taken to imply that there was reason not to use the simple factive form).

In reconstructing the processes behind the other's action, inverse planning locates the agent's reason or reasons for acting, as far as possible, within a shared world of facts; and likewise, as I discuss in the final section, what the agent's emotions are about, Shared world explanations have a number of advantages over those requiring explicit mentalizing: they can identify environmental threats and rewards, they are conceptually and linguistically less demanding, and they achieve greater code compression. If this is correct, then we must reject the common assumption that explicit mentalizing, or mental state attribution, is the paramount explanatory aim of the procedures we lump under the term *mentalizing*. The aim is rather to interpret behavior in terms of a shared world where this is possible and to diagnose cases where it is not.

We can of course add to any theory a stipulation that the interpretation starts by importing the world of the interpreter. Rebecca Saxe, a leading proponent of the theory-based approach to mentalizing in neuroscience, writes:

> I agree that by far the bulk of action explanation in every day life is accomplished by 'factive', 'agent-neutral' coding of beliefs (and indeed of desires!). When I try to explain this, I sometimes talk about the default naive realism we bring to understanding both the world and other people. Instead of beliefs or perceptions, we explain actions in terms of facts. Instead of desires, we explain actions in terms of what is valuable or good. Explanations in terms of mental states (what she saw, or didn't see, or thought, or wanted) are exceptions, corrections. (Personal communication, July 6, 2020)

It should be remarked that agent-neutral coding requires a simulative account of inverse planning. It stipulates that top-down inputs are by

default invariant between the direct, or forward, employment of the action planning system and its inverse simulative use in interpreting another's behavior. If Saxe indeed accepts the simulative account of inverse planning, with its reuse of the very system used for planning and generating one's own actions, all well and good: what are facts for us are portrayed as available to others' decision making as well – and therefore, as I will argue, as something known to the others. If on the other hand it is simply plastered onto a formal theory or model, perhaps as a useful heuristic, then we can't speak of an automatic carry-over of an agent-neutral (same for self and other) coding.

Agent-neutral coding requires the simulation account of inverse planning, with its concurrent use of the same system for generating actions and interpreting the action of others; and, as I will argue, it is agent-neutral coding that explains why what we ourselves regard as *facts* get passed along to other (the target agent) as *known* facts. However, not all versions of the simulation theory of mentalizing support default agent-neutral coding. The simulation theory has sometimes been characterized as a two- or three-step process of first reading one's own mental states (by introspection or otherwise) and then inferring that the other agent has similar mental states. Many proponents, as well as most critics of the simulation theory, have supposed simulation to be founded on such an implicit inference from oneself to others. The form of inference is essentially the old argument from analogy (Mill 1869), which requires that one first recognize one's own mental states under actual or imagined conditions and then infer that the other is in similar states. This is usually linked to an introspectionist account of how one recognizes and ascribes one's own mental states (Goldman 1993). It is further assumed that, to recognize and ascribe one's own mental states and to mentally transfer these states over to the other, one would need to be equipped with the concepts of the various mental states. According to this account, in short, simulation is an analogical inference from oneself to others premised on introspectively based ascriptions of mental states to oneself, requiring prior possession of the concepts of the mental states ascribed. Goldman's account of simulation has been characterized as requiring three stages of processing in order to generate an interpretation of another's behavior:

> Stage 1. *Mental simulation*: Subject *S* undergoes a simulation process, which outputs a token simulated mental state *m**.
> Stage 2. *Introspection*: *S* introspects *m** and categorizes/conceptualizes it *as* (a state of type) *M*. (Barlassina, Gordon 2017)
> Stage 3. *Judgment*: *S* attributes (a state of type) *M* to another subject, *Q*, through the judgment *Q is in M*.

In short, we (or our brain) must somehow read our own mental states, then describe or categorize them, and finally form a judgment that the other is in the same or similar state. However, given the simple alternative of agent-neutral coding, with one and the same neural code indifferently serving both self and other, this elaborate intellectually loaded process seems both unnecessary and wasteful of time as well as of energy resources.

5 Perspective-Taking and Positional Correction

The most economical strategy for mentalizing, other things being equal, would be one that minimizes individuation, or information tagged to specific individuals. That is, it would minimize the need for explicit mentalizing, in the sense of judgments about mental states or processes. In the default case, with uncorrected agent-neutral coding, the actions of others would be interpreted in terms of a shared world – that is, to the world on the basis of which we ourselves act. Mentalizing, on this account, would be called on to complement or to correct what is passed along through agent-neutral coding. It would be reserved for cases in which a shared world proves inadequate to predict or explain the actions or emotions of particular individuals.

Spatial perspective-taking is probably the most familiar type of error correction in the interpretation of others' behavior. Moving mentally to the other's viewpoint, we may recognize that their view is partially or wholly occluded (they are in a different room), or we recognize that they can see aspects of a scene that are hidden to us, and consequently that they may know something we do not know. As Nagel writes,

> the capacity to differentiate patterns of knowledge and ignorance in our fellow agents enables us to exploit their epistemic access to those parts of reality for which their vantage point is better than ours. If you want to know which way the coin in my palm is facing, you know you can ask me. While many primates show selective social learning from peers recognized as knowledgeable, humans show exceptionally active use of the knowledge of their peers (Tomasello 2019), guided by an exceptionally well-developed sense of what others do and do not know, a sense informed by continual feedback from conversational exchanges (Westra and Nagel 2021) and extraconversational encounters with reality. (Nagel 2023, 206)

In addition, rather than imparting different information, the altered viewpoint may account for a different emotional or motivational response. To a stranger observing the scene from a distance, the bear now approaching me is not likely to feel threatening, or in any case

as threatening as it does to me. The threatening (or non-threatening) emotive quality of the bear may be seen as a function of one's location relative to the bear – or, the bear's location and vector in egocentric space. With the ability to move mentally into another's spatial perspective, individual differences become mere positional differences. That is, it is a good starting bet that (unless there is evidence to the contrary) any individual in the same position will see the bear as threatening. With the operation of 'putting ourselves in the other's place' by spatial perspective-taking, we are able to restore the economic advantages of a shared world. We allow the threatening quality to remain out there in the bear, or rather in the bear from a point of view. We need not represent it as a function of individual mental makeup, even if some individuals may be found immune to the standard bear-approaching-me response.

Although it is spatial perspective-taking that gives us the general metaphor of 'perspective-taking,' 'adopting the other's point of view,' and 'putting ourselves in the other's place,' many other kinds of corrections may be considered broadly perspectival, or positional. For example, differences in social or occupational role may be bridged by a kind of perspective shift: student/teacher, worker/manager, diner/waiter, patient/doctor, consumer/salesperson. In these cases, as in differences in spatial perspective, it may be sufficient to shift to a generic 'point of view,' or, as we say, to understand where the other is 'coming from,' to explain the other's actions, without explicit mentalizing. That is, it may be a good starting assumption that a person in a given 'position' will act in more or less the same 'standard' way, an assumption that may underlie the notion of generic 'scripts' of action sequences postulated by Schank and Abelson (1977). Such an assumption would exploit positional redundancies and limit new input to deviations from the standard.

6 An Evolutionary Perspective

For most of human history, social encounters would have occurred primarily within small, close-knit cultural groups with limited exposure to faraway lands and diverse cultures. As a result, to explain and predict behavior within the local group, 'mentalizing' could have consisted largely of looking to the shared world and its common facts, emotive qualities, affordances, attractions, and repulsions.. The environmental and cultural contexts of these small social groups led to the development of shared mental maps and a common understanding of their surroundings. Members of these groups would have agreed on which elements of their environment were significant, threatening, appealing, or repulsive. This shared understanding allowed for relatively straightforward predictions and explanations of

each other's behavior, given the group's limitations and homogeneity. Of course, even in these close-knit communities, individual differences in temperament, sensory and cognitive capacities, knowledge, acculturation, and goals existed. However, such differences would have been relatively rare and likely observed against the backdrop of the more predictable shared background. In such situations, minor adjustments could be made to accommodate these individual differences.

The evolutionary advantage of this social predictive system lies in its ability to exploit, reinforce, and create redundancies within the group. The more shared understanding and predictability there is among group members, the smoother the social interactions and cooperation, leading to increased chances of survival and successful reproduction.

The process of social learning and prediction plays a vital role in fostering unity and cooperation within small groups. Infants and young children acquire knowledge by observing and imitating the behavior of trusted adult caregivers. Through social referencing, they learn how to react to various situations and stimuli based on the responses of those they trust. By imitating similar responses, the child's behavior aligns with the group's norms and expectations, leading to shared patterns of behavior that are strengthened and repeated.

However, as societies evolved and expanded, encounters with culturally distant and geographically separated groups became more frequent. In such encounters, the strategy of agent-neutral coding that worked reasonably well within small, homogenous groups might no longer be effective. Understanding and predicting the behavior of people from vastly different cultures would require extensive corrections and adjustments, as their mental maps, norms, and affordances could vary significantly from one's own.

In summary, the evolutionary perspective suggests that the reliance on agent-neutral coding and shared mental maps was an effective strategy for understanding and predicting behavior within small, culturally cohesive groups. However, as human societies became more complex and interconnected, this strategy faced limitations in explaining behavior in culturally distant contexts, necessitating the development of more nuanced and culturally sensitive approaches to cross-cultural understanding.

7 Ignorance and False Belief

How does inverse planning deal with ignorance? For example, we see someone do something surprising: in broad daylight, they walk nonchalantly into a deep puddle. We are aware of the puddle, but apparently the other, engrossed in their cellphone, is not: Earlier, I cited Nagel on knowledge recognition. For facts automatically passed

along by agent-neutral coding, perhaps the more important capacity is *ignorance* recognition. We pick up on evidence of behavior that is *not* truth-anchored, and accordingly, we modify the default input to inverse planning. We make the surprising behavior unsurprising by disconnecting or 'decoupling' the fact that there was a puddle in his path from the input to inverse planning. Decoupling a fact from inverse planning is a way of marking ignorance of a fact. Ignorance, in turn, may engender false belief because the puddle-walker was ignorant of the fact that there was a puddle. Out of touch with the facts concerning his current environment, they continued operating on the false default assumption of an ordinary puddle-free path. The puddle is there, but it is not there for the other – until it is.

Agent-neutral coding and the possibility of toggling between knowledge and ignorance would give us the neural underpinnings for two theses long held by the psychologist Josef Perner: first, that well before they have an explicit grasp of belief attribution, young children are quite capable of explaining action in terms of the external situation; and second, that older children and adults use the same type of explanation young children use, except in the occasional cases where it proves inadequate; then they must fall back on explanations that mention the mental states, especially the beliefs, of the agent. Young children and, where possible, older children and adults

> make sense of intentional actions in terms of justifying reasons provided by 'worldly' facts (not by mental states). (Roessler, Perner 2013, 35)

The young child's conception is all we usually call upon, because it is typically all we need. This comes to saying that explaining and predicting actions in terms of actual situations or facts is our default mode of explanation and prediction, the mode we employ unless we find some reason not to. Only where this appears inadequate do we invoke beliefs in our explanation. In the classic 'false belief' condition, you see individual A place her treasure at location *x*. You also see that (*m*) the treasure has been moved and is now at a different location *y*.

If you were planning to steal the treasure, your action planning system would take account of (*m*) and direct you to location *y*. However, if your system is hypothetically generating A's plan to retrieve A's treasure, the question arises: Does A know about the move? Is A aware that (*m*)? The possibility of attributing ignorance, or not knowing, is simply the possibility of decoupling the action planning system from the fact that (*m*). (*Egocentric* ignorance acknowledges that there are facts to which our own planning is not yet coupled or connected.) *Knowledge*, on the other hand, is represented simply by nonintervention. That is, one implicitly attributes knowledge that (*m*) simply by *not decoupling* the system from the fact that (*m*). 'Knowledge

representations' accordingly consist in nothing more than *access to facts*.

Attributing ignorance consists in decoupling from fact, which is an extra step beyond implicitly attributing knowledge. False belief requires decoupling as well as introducing into the planning process an 'as if' fact, such as that the treasure is still at location *x*. True belief for the wrong reason would similarly entail introducing an 'as if' fact. (Although it might produce the same actions as the 'real' fact, the counterfactual dependencies would differ.) The upshot is that what is really basic is a shared world, where, prior to any corrective processing, everything we ourselves regard as the world, as the facts, is publicly accessible and thus available to others as possible reasons for action.

8 Knowledge First

It is traditional to see factual knowledge as an achievement, as having a status that is to be earned by meeting certain stringent conditions. As Nagel suggests, we develop the capacity to recognize when those conditions indicate a state of mind of a type that one can only have to truths. Consistent with this, however, is that knowledge is also a status granted by birthright, as it were. When we try to make sense of another's actions and emotions, we gift the other with access for planning – and, I will argue, for emotion generation – to all the facts available to us in generating our own actions and emotions: that is, with knowledge of these facts. (There may be differences in attention, of course: for one thing, our own direction of gaze may differ from the other's. But this is often a bridgeable gap: we look around to previously unnoticed features of the environment, or more broadly, to aspects of the world that might be salient to the other.

In summary, there is evidence that the human brain exploits a strategy that appears to operate in several other areas of cognition, that of analysis by synthesis. Specifically, the brain interprets the behavior of others by testing hypothetical ways of *generating* that behavior. This would involve the inverse use of one's own system for planning and generating intentional action, concurrent with its primary 'forward' use in generating one's own actions. The inverse use of the planning system for hypothetically generating the actions of others would ordinarily require adjustments of the top-down inputs to the system. These would include adjustments of the factual input, the set of facts that influence planning. In hypothetically generating another's actions, the planning system may be selectively decoupled (disconnected, unplugged) from some of these facts. In a predictive strategy, the actual world – that is, what we ourselves take to be the facts – serves as a starting point, an opening bid or bet, subject to

revision ('correction') on the basis of new evidence. As our mechanisms for decision-making and planning are used to test hypothetical explanations of the actions of others, the carryover of agent-neutral inputs has the effect of projecting onto others a shared world within which we act and interact. Strictly speaking, the brain doesn't do anything to accomplish this; rather, it is by not doing anything to modify or correct the top-down inputs in the concurrent use and reuse of action planning that gives us a shared world as a default. Withholding or diminishing the implicit attribution of knowledge, such as attributing a belief that falls short of knowledge, requires additional steps in neural coding and processing. Those extra steps, their added complexity and their drain on resources, suffice to explain the empirical findings: why (per Phillips 2021) some individuals – non-human primates, young children, and certain cognitively impaired people – can attribute knowledge but not belief, while none attribute belief but not knowledge; and why attributions of knowledge are 'more automatic' than those that require additional processing.

9 Emotions

Emotions are not, in any straightforward sense, planned and executed as actions. They may sometimes be expressed in action, as in 'acting out of anger'; it is also possible to allow oneself to be angry, as well as to decide to interpret as anger interoceptive responses that are ambiguous. But in general, anger is not generated by action planning, and its interpretation by others is therefore not a function of *inverse* planning. However, it is plausible that the processes responsible for emotion generation, whatever their nature, can be interpreted by inverting them. In inverse emotion generation, we test hypothetical ways of generating something approximating the emotion we observe.

Suppose we see someone look startled, or frightened, or obviously pleased about something, but we can't easily tell the source of the emotion. Following the other's gaze, we find several objects or environmental features, any one or more of which might be the source: we need to pick out the *right* feature or features. Or suppose the person has already turned away from the source of the emotion. What do we do in such cases? We look around for a plausible target. That is, we look for something startling. Or if the other is frightened, we look for something that is frightening. If pleased, we look for something pleasing. To do this, we engage *our own* system for generating emotions out of our perceptions. We are also prepared to make positional adjustments, where necessary. For example, I see a competitor for an award and find her looking elated. On the wall nearby there is posted a list of award-winners. My own name on the list

would indeed be pleasing; but I automatically shift to viewing the list through her eyes. We do this sort of thing so routinely that we aren't aware of doing it – and we fail to appreciate the sophistication of the maneuver we are engaging in.

10 Conclusion

If we understand simulation in terms of default agent-neutral coding, then we have to reject a well-known account of the simulation theory: that it requires introspective recognition of one's own (actual or pretend) mental states (metacognition), followed by attribution of the same states to the other individual (Goldman 2006). Agent-neutral coding clearly would support a more economical account of simulation, one that requires neither metacognition nor self-other inference (Gordon 1995). It is simply by default that the inputs to inverse planning are the same as the inputs to forward self-planning; This carryover is not established through an inferential leap from self to other, but rather, as I suggested, simply by omission: that is, crossing the self-other border without doing anything to *alter* the existing inputs.

Bibliography

Baker, C.L.; Saxe, R.; Tenenbaum, J.B. (2009). "Action Understanding as Inverse Planning". *Cognition*, 113, 329-49.

Baker, C.L.; Saxe, R.; Tenembaum, J.B. (2011). "Bayesian Theory of Mind: Modeling Joint Belief-Desire Attribution". *Proceedings of the Cognitive Science Society*, 32, 2469-74.

Baker, C.L.; Tenenbaum, J.B.; Saxe, R. (2006). "Bayesian models of human action understanding". *Advances in Neural Information Processing Systems*, 18, 99-106.

Barlassina, L.; Gordon, R.M. (2017). "Folk Psychology: as Mental Simulation". Zalta, N. (ed.), *The Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/archives/sum2017/entries/folkpsych-simulation/

Blakemore, S-J.; Frith, C. (2005). "The Role of Motor Contagion in the Prediction of Action". *Neuropsychologia*, 43, 260-7.

Bouquet, C.A. et al. (2011). "Motor Contagion: Goal-Directed Actions Are More Contagious Than Non-Goal-Directed Actions", *Experimental Psychology*, 58(1), 71-8.

Clark, A. (2013). "Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science", *Behavioral and Brain Sciences*, 36(3), 181-204.

Dietz, C.H. (2018). "Reasons and Factive Emotions". *Philosophical Studies*, 175, 1681-91.

Goldman, A.I. (1989). "Interpretation Psychologized". *Mind and Language*, 4(3), 161-85.

Goldman, A.I. (1993). "Consciousness, Folk Psychology, and Cognitive Science". *Consciousness and Cognition*, 2(4), 364-82.

Gordon, R.M. (1969). "Emotions and Knowledge". *Journal of Philosophy*, 66(13), 408-13.

Gordon, R.M. (1986). "Folk Psychology as Simulation". *Mind and Language*, 1, 158-71.

Gordon, R.M. (1987). *The Structure of Emotions: Investigations in Cognitive Philosophy.* Cambridge, UK: Cambridge University Press.

Gordon, R.M. (1992). "The Simulation Theory: Objections and Misconceptions". *Mind and Language*, 7(1-2), 11-34.

Gordon, R.M. (1995). "Simulation Without Introspection or Inference From Me to You". Davies, M.; Stone, T. (eds), *Mental Simulation: Evaluations and Applications*. Oxford: Blackwell, 53-67.

Gordon, R.M. (2000). "Sellars Rylean Revisited". *Protosoziologie*, 14, 102-14.

Gordon, R.M. (2021). "Simulation, Predictive Coding, and the Shared World". Gilead, M.; Ochsner, K.N. (eds), *The Neural Basis of Mentalizing*. Dordrecth: Springer, 237-56.

Heal, J. (1986). "Replications and Functionalism". Butterfield, J. (ed.), *Language, Mind, and Logic.* Cambridge: Cambridge University Press, 45-59.

Nagel, J. (2023). "Seeking Safety in Knowledge". *Proceedings and Addresses of the American Philosophical Association*, 97, 186-214.

Phillips, J. et al. (2021). "Knowledge Before Belief". *Behavioral and Brain Sciences*, 44, 1-37.

Roessler, J.; Perner, J. (2013). "Teleology: Belief as Perspective". Baron-Cohen, S.; Lombardo, M.; Tager-Flusberg, H. (eds), *Understanding Other Minds: Perspectives from Developmental Social Neuroscience, 3rd edition.* Oxford: Oxford Academic Press, 35-50.

Schank, R.C.; Abelson, R.P. (1977). *Scripts, Plans, Goals, and Understanding. An Inquiry into Human Knowledge Structures*. New York: Psychology Press.

Thalberg, I. (1964). "Emotion and Thought". *American Philosophical Quarterly*, 1, 45-55.

Williamson, T. (2000). *Knowledge and Its Limits*. Oxford: Oxford University Press.

Yildirim, I.; Siegel, M.; Tenenbaum, J.B. (2020). "Physical Object Representations for Perception and Cognition". Poeppel, D.; Mangun, G.R.; Gazzaniga, M.S. (eds), *The Cognitive Neurosciences, 6th edition.* Cambridge: The Mit Press, 399-410.