Simulation Theory

Intermediate article

Joe Cruz, Williams College, Williamstown, Massachusetts, USA
Robert M Gordon, University of Missouri, St Louis, Missouri, USA

CONTENTS

The simulation theory is an account of our everyday ability to attribute mental states and predict and explain human behavior. It has been developed both as an empirical hypothesis in cognitive science and as an account of mental concepts in the philosophy of mind.

WHAT IS THE SIMULATION THEORY?

The simulation theory (ST) is an account of our everyday ability to make sense of the behavior of others. One crucial element of this ability is the identification and attribution of inner mental states that generate action, especially propositional attitudes such as beliefs or desires. The successful 'mindreading' of mental states allows us to predict and to explain what others do, and makes possible the rich social dynamic that pervades human life.

Conceived most broadly, ST maintains that one represents the mental activities and processes of others by mental simulation, i.e., by generating similar activities and processes in oneself. For example, one anticipates the product of another's theoretical or practical inferences from given premises by making inferences from the same premises oneself. In more complex simulations, one imaginatively adopts the circumstances of the target and then uses one's own mental apparatus to generate mental states and decisions. Computationally, this exercise of imagination is usually represented as feeding pretend inputs into one's own decision-making processes, taking these processes 'offline' so that they do not issue forth in real behaviors.

Some proponents of ST go further and claim that many of the concepts of mental states that we deploy in understanding other human beings are fundamentally linked to our possession of those same mental states. Some prominent accounts attempt to shed light on the conceptual transformation involved in refashioning our first-person concepts in such a way that they can be deployed in the third person.

While ST is related to the empathetic or *verstehen* approaches to explanation in the social sciences that were prominent in the twentieth century, most researchers who currently work on ST do not operate directly within that theoretical framework. In its contemporary forms, ST has often been developed with its principal rival, the theory theory (TT), as an explicit foil. TT maintains that the mental terms and concepts used in understanding, predicting and explaining human behavior derive from a folk theory of the mind. A closer historical source for contemporary ST was the debate in the philosophy of mind over the status of this putative theory. According to one view, known as eliminativism: (1) mental states like beliefs and desires are the posits of a folk theory of the mind; and (2) this theory is radically false. The conclusion drawn by the eliminativist is that the faulty folk theory ought to be rejected in favour of some more scientifically respectable theory such as one derived from neuroscience.

Before the advent of ST, most critics of eliminativism focused on the defensibility of its second tenet. Important articles by Jane Heal (1986), Robert Gordon (1986) and Alvin Goldman (1989) challenged the first tenet by setting out the ST alternative to TT. If it could be shown that mental terms did not derive their intelligibility from their role in a folk theory, then the eliminativist's conclusion would become suspect. Since this important impetus, ST has attracted interest in its own right and has become somewhat independent of concerns over eliminativism.

Some philosophers believe that ST sheds light on traditional topics such as the problem of other minds, referential opacity, broad and narrow content, and the peculiarities of self-knowledge. ST has had a substantial impact on research into 'theory of mind' in developmental psychology, as well as on branches of philosophy outside the philosophy of mind, especially aesthetics and the philosophy of the social sciences.

SIMULATION THEORY VERSUS THEORY THEORY

One of the dominant explanatory patterns within cognitive science and the philosophy of mind has been to construe a mental capacity as subsumed by a theory of the domain of the capacity. The idea is that domains such as folk physics, folk biology, and intuitive statistics are treated as areas of knowledge in which the layperson's judgments are the results of applying a theory. The results of the application of the theory constitute our spontaneous conscious judgments about cases that seem amenable to the theory. Such theories are usually thought to be representations of law-like generalizations. In some domains, at least, they may be characterized as a set of 'platitudes' tacitly deployed in thinking about problems and circumstances in our world. The layperson need not be aware of using a theory to make a judgment, but, presumably, could generate at least some of the platitudes that inform his or her judgments.

In keeping with this widely endorsed strategy in cognitive science, theory theorists claim that we possess a body of tacit knowledge that governs our judgments about the mental states of others. The theoretical posits of this theory will be mental states like beliefs and desires, and the transitions between the mental states will be described by the theory as mental processes such as inference.

There is a diversity of opinion among theory theorists on the nature and origin of the putative common-sense theory. Some claim that the theory is learned; others that it is innate. Theory theorists also face a question regarding the modularity of the common-sense 'theory of mind'. Some claim that the theory is a distinct cognitive module; others that it is continuous with the system of representations that constitute theories of other, non-mental domains. What unifies theory theorists is the view that attributing inner states and making sense of the behavior of others is carried out by a capacity that deploys knowledge encoded in a theory.

The most straightforward sense in which ST is opposed to TT is that simulation theorists deny that our capacity to attribute mental states is subsumed by a body of knowledge about the minds of others. Rather, our own mental processes are treated as a manipulable model of other minds. Such simulation would typically require indexical adjustments, such as shifts in spatial, temporal, and personal points of view, to place oneself in the other's physical and epistemic situation in so far as it differs from one's own. One may also compensate for the other's reasoning capacity and level of expertise, if possible, or modify one's character and outlook as an actor might, to fit the other's background and behavioral history. With these adjustments, the attributer might enter mental states that differ from those he or she would have in the target's situation. Even when simulation is insufficient for making decisions in the role of the other, it might allow one to discriminate between those options likely to be attractive to the target and those likely to be unattractive. Accordingly one would be prepared for the former actions and surprised by the latter.

Moreover, most simulationists are happy to grant that, in some cases, we will develop general rules of thumb or heuristics for attributing mental states. These may be called on to generalize the results of a simulation to cover the same target at future times, or a class of targets, such as those who share the conventions of a particular culture, who 'as a rule' behave in a certain way in certain circumstances. Still, simulationists deny that general, 'theoretical' considerations play a fundamental part in attributing mental states. ST is often characterized as 'process driven', because it is a cognitive process that is generating the output of the simulation, with little or no influence from general information about minds.

An analogy may be helpful here. If we wished to predict a future state of the solar system, we might appeal to a theory that expresses law-like generalizations about the motions of the planets. So, we might appeal to the theories articulated in a contemporary textbook on astronomy, or, more in keeping with our interest here in folk theories, we might appeal to a set of platitudes about the motions of celestial objects of the kind articulated by the ancients. This would be a sensible approach, but it is not the only way we could successfully carry out the prediction. If we could build a reasonably accurate physical model (an orrery), we could advance this model the correct number of cycles and read from it a future state of the solar system. Depending on how versatile our model was, we might even be able to experiment with counterfactual starting states; or, with a still more sophisticated model, or a digital simulation, we might even adjust the orbits of planets, and their number, to model a range of planetary systems quite unlike our own.

This analogy gives us some insight into the difference between theory-driven and process-driven accounts of a domain. ST can be viewed as the proposal that we use our own mental states and processes – our perceptual, cognitive, motivational and emotional systems – as a model like an orrery.

VERSIONS OF SIMULATION THEORY

It is obviously desirable that any relevant disparities between simulator and target should be removed or offset in some way. If the behavior to be predicted or explained is crucially dependent on beliefs, desires or emotions not shared by the simulator, then the simulator must either adopt the appropriate pretend-beliefs, pretend-desires or pretend-emotions, or compensate by a heuristic rule. However, this proviso says nothing about the nature of belief, desire or emotion. For instance, it is consistent with (but certainly does not imply) a functionalist account of belief. In a well-known box diagram, Stich and Nichols (1992) portray ST as the empirical hypothesis that the same belief–desire system that generates one's own decisions and actions also generates our predictions of the decisions and actions of others, adding for this purpose a pretend-belief generator and a pretend-desire generator. Some simulationists find this portrayal too restrictive, claiming that it commits ST to a questionable conception of mental states and possibly also a mistaken understanding of the dependence of actions on these states. These proponents of ST conceive our everyday ascriptions of belief and other mental states as part of an explanatory enterprise quite unlike the attempt to fill in the 'boxes' of a functional theory like those commonly developed in cognitive science. Our ascriptions specify the 'internal states' of a system only in the sense of attempting an essentially first-person glimpse into a subject looking out on the world.

There is also disagreement among simulationists on another front. Some hold that to ascribe mental states to others by simulation, one must already be able to ascribe mental states to oneself by introspection, and that to do this one must already possess the relevant mental state concepts. On this view, simulation is understood as essentially an application of the argument from analogy. Others attempt to build on the 'subject looking out on the world' idea of mental state ascription. They hold that in such ascriptions, whether concerning oneself or another, one is saying something about the world, albeit in a way that is relativized to a particular 'point of view'. Rather than resting on an analogy between what lies 'inside' two individuals, this account assumes that, unless there is evidence to the contrary, all subjects look out on one and the same world.

ARGUMENTS FOR SIMULATION THEORY

A number of arguments have been put forward in favour of the simulation theory, three of which are outlined below.

Parsimony

The most important distinguishing feature of folk psychology, according to many writers, is the central and essential role it gives to the semantic content of the states it posits, particularly the propositional or sentential 'objects' of propositional attitudes such as beliefs, desires, and intentions. Most theory theorists try to accommodate this feature with the hypothesis that folk psychology comprises laws or principles that quantify over this content, connecting, for example, what someone believes and desires to what that person chooses to do. Moreover, the connections are said generally to mirror the semantic relations that hold among these contents, particularly relations that can be represented abstractly by rules of logic and rational argument such as modus ponens and the practical syllogism. Thus the theory theory posits an internal store of causal laws corresponding to these rules.

However, in so far as the store of causal generalizations mirrors the set of rules to which our own thinking typically conforms, ST appears to render it otiose. For whatever those rules are, our thinking continues to conform to them within the context of simulation, unless, of course, adjustments are made to accommodate evident differences. In short, we can use our own reasoning as a model of the reasoning of beings that reason the way we do. In the light of this alternative, it is argued, the hypothesis that people must be endowed with a special stock of laws corresponding to rules of logic and reason appears unmotivated and unparsimonious.

Other Uses of Simulation

The procedure that ST posits as essential to the common-sense methodology for predicting and explaining behavior also appears to be – with modifications – essential to emotional empathy, and important if not essential to ethical evaluation. Even if one didn't think simulation essential to common-sense explanation and prediction, one would probably have to posit such a procedure anyway to account for empathy and ethical evaluation.

Explanation of Children's Errors in Predicting Behavior in 'False Belief' Situations

According to ST, the mature capacity for explaining and predicting the behavior of others requires capacities for imaginative pretense of at least three kinds: counterindexical pretending, which recenters the egocentric map (I am spatially or temporally somewhere else in the world, or I am someone else); counterfactual or propositional pretending, in which the world itself is altered in imagination (for example, this banana is a telephone, or dinosaurs roam the boulevards of Paris); and what might be called purposive pretending, in which alternative goals are adopted (for example, the putative goals of a mother, or the goals of an opponent in a game). These capacities are clearly evident in most children before their third birthday, and they are typically combined in role play.

To explain and predict a great deal of the human behavior they are likely to encounter in real life or in stories, young children can probably get by with a relatively simple employment of these and similar imaginative abilities. For example: while Sally and Anne are playing together, Anne grabs the marble in Sally's box and places it in her basket. Taking the role of Sally in this simple scenario, a child should have no trouble deciding where to look for her marble: in Anne's basket. However, suppose that in the story (or in a scene witnessed in real life) Sally is away when Anne takes her marble. When she returns, where does she look for her marble? Taking the role of Sally and deciding where to look, the child would be misled by a simple use of pretense. To get the correct prediction (she will look in her box, where she left it), and at the same time to maintain 'objectivity' (for example, to recognize that she won't actually find the marble there), the child would have to feed contrary premises into two distinct lines of reasoning: 'objectively', the marble has been moved to Anne's basket and therefore can be found there, not in Sally's box; 'subjectively' – and for the purpose of deciding what to do in the role of Sally – it hasn't been moved from Sally's box, and therefore can still be found there. According to ST, until children are capable of such compartmentalized reasoning – and of knowing when to use it – they can be expected to make incorrect predictions in complex scenarios in which behavior is likely to be based on a false belief.

Numerous experimental studies (beginning with Wimmer and Perner (1983)) have confirmed that in fact children do generally make these incorrect predictions until about the age of four. Although ST is not unique in offering an explanation of such errors, its explanation appears more compelling than typical 'theory' explanations, such as the hypothesis that children lack the 'belief' part of the theory of mind until the age of four (Gopnik, 1993; Wellman, 1990; Gopnik and Meltzoff, 1997).

SIMULATION THEORY AND COGNITIVE SCIENCE

The simulation theory has a bearing on a range of disciplines and methods in the cognitive sciences. Consequently, there has been significant research on ST within artificial intelligence (AI), neuroscience, and cognitive psychology. In order to show how ST has been pursued in these fields, it is useful to develop further the 'process driven' conception of ST introduced earlier. One way of refining the claim that simulations are process-driven is to take ST as the hypothesis that essentially the same set of mechanisms is called upon to provide two different competences. One of these competences is the intelligent control of behavior. This would include, among other things, the capacity to make inferences from beliefs to new beliefs and the capacity to make decisions on the basis of beliefs. The second competence is the anticipation and comprehension of intelligently controlled behavior, by predicting the underlying inferential and decision-making processes.

There are thought to be testable consequences of this 'double-duty mechanism' construal of ST. Researchers in AI have claimed that the same models that provide an account of practical reasoning (and of other dynamics between propositional attitudes) can efficiently be adapted to perform mental state attribution. These computational efficiency arguments have in turn sometimes been used as arguments in favour of ST. Both traditional AI programming methods and neural network techniques have been employed in research on ST along these lines.

In AI research, the double-duty hypothesis is a functional claim. The idea is that the same mental program is implicated both in practical reasoning and in mental state attribution, with changes allowed for taking the system 'offline' or feeding in pretend states. So, even if it turned out that different parts of the physical computational system were responsible for practical reasoning and mental state attribution, the AI approach to ST would remain valid.

If, on the other hand, the double-duty conception of ST is not just understood functionally, but includes a commitment to the commonality of the underlying neural substrate, we arrive at another avenue of cognitive science research on ST. Such research would seek to show that there was a shared neuronal mechanism that is responsible for practical reasoning and mental state attribution. This approach is analogous to a fruitful line of research on vision and visual imagery. These two capacities appear to share substantial portions of the underlying neural substrate.

There is emerging evidence for the existence of 'mirror neurons' in humans and other primates (Fadiga *et al*., 1995; Rizzolatti *et al*., 1996). Single-cell recordings in macaque monkeys and magnetic stimulation techniques in humans show that these mirror neurons exhibit increased activity when another primate is observed performing some characteristic action, such as grasping an object. This is relevant for ST because these are the same neurons that show increased activity in the first-person performance of that action. This research is far from showing that sophisticated cognition is subsumed by double-duty mechanisms, but it is suggestive. Some simulation theorists maintain that this research reveals an evolutionary precursor to the capacity of assuming a different perspective that ST requires (Gallese and Goldman, 1998).

Evolutionary considerations have been lurking in the background of much empirical research in this domain. Thus, elements of cognitive ethology have been thought to bear on ST, and vice versa. Indeed, it was in their research on primate behavior that Premack and Woodruff (1978) introduced the term 'theory of mind'. Though efforts to develop ST within cognitive ethology have been limited, it does present another potential source of data.

A POSSIBLE TEST OF THE SIMULATION THEORY

If it can be shown that incorrect mental state attributions are in some cases best explained by claiming that the attributer lacks some specific and perhaps surprising background information about mental agents, then it seems less plausible to think that the attributer is simulating. Some critics of ST claim to have uncovered experimental evidence for just this.

The logic of such experiments is thought to be as follows. In order for a simulation to be successful, the operation of the attributer's mental processes must be substantially similar to the operation of the target's mental processes. The attributer does not need to know about the vagaries of human psychology because, by hypothesis, his or her own mental mechanism is subject to those same vagaries. One test of ST, according to this line of reasoning, would be to select some surprising feature of our mental life to see whether subjects can successfully attribute mental states to others where the mental processes in question exploit the surprising feature. If the attribution fails, this suggests that either the attributer's mind does not share the surprising feature, or that it is the specific lack of information about the surprising feature that is generating the failure to attribute a correct mental state. Neither of these results would be friendly to ST. The first possibility is against the very spirit of ST, while the second possibility implicates a theory of mental states.

In one attempt to pursue this line of criticism, researchers explored the counterintuitive effect (reported by Langer (1975)) whereby subjects demand more money in exchange for a lottery ticket that they have chosen than for a ticket that has just been given to them. Although there has been some concern about the reproducibility of Langer's initial results, the experiment is thought to show something unexpected about human psychology, namely, that subjects are more attached to items that they have chosen than they are to identical items that they have not chosen. Nichols *et al*. (1996) asked naive subjects to predict the outcome of the Langer experiment without actually putting them under the experimental conditions themselves. It was assumed that if, in making the prediction, the subjects simulated making a decision themselves under each of the two experimental conditions, they would be prone to the same surprising effect, valuing the ticket they had chosen themselves more than the one they had not. Thus, they would predict correctly. However, the subjects did not predict correctly, and Nichols *et al*. concluded that they were not simulating. Several problems have been noted with the experiment conducted by Nichols *et al*. A more discriminating set of experiments has been reported, with mixed results for ST (Perner *et al*., 1999).

One general objection concerns the logic of such tests. The account presented above neglects the following possibility: the experimental effect in question, such as the higher value placed on items one has chosen, may be partially due to aspects of processing that merely imagining does not, or perhaps even cannot, capture. Consider the two lines in the Mueller–Lyer illusion: remove the arrowheads, and you are likely to judge the lines to be of equal length; merely *imagine* the arrowheads removed, and you are still likely to judge them unequal. Analogously, one may predict what people will do in situation S by simulation, in the sense of imagining being in S and deciding what to do, and yet fail to replicate all the processing that would occur if one actually were in S. If the analogy holds, then tests of ST should take account of a possible gap between imagining and replicating at a subpersonal level.

References

Fadiga L, Fogassi L, Pavesi G and Rizzolatti G (1995) Motor facilitation during action observation: a magnetic stimulation study. *Journal of Neurophysiology* **73**: 2608–2611.

Gallese V and Goldman A (1998) Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences* **2**: 493–501.

Goldman A (1989) Interpretation psychologized. *Mind and Language* **4**: 104–119.

Gopnik A (1993) How we know our minds: the illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences* **16**: 1–14.

Gopnik A and Meltzoff A (1997) *Words, Thoughts and Theories*. Cambridge, MA: MIT Press.

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

Heal J (1986) Replication and functionalism. In: Butterfield J (ed.) *Language, Mind and Logic*, pp. 135–150. Cambridge, UK: Cambridge University Press.

Langer E (1975) The illusion of control. *Journal of Personality and Social Psychology* **32**: 311–328.

Nichols S, Stich S, Leslie A and Klein D (1996) Varieties of off-line simulation. In: Carruthers D and Smith E (eds) *Theories of Theories of Mind*, pp. 39–74. Cambridge, UK: Cambridge University Press.

Perner J, Gschaider A, Kühberger A and Schrofner S (1999) Predicting others through simulation or by theory? A method to decide. *Mind and Language* **14**: 57–79.

Premack D and Woodruff G (1978) Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* **4**: 515–526.

Rizzolatti G, Fadiga L, Matelli M *et al*. (1996) Localization of grasp representations in humans by PET: 1. Observation vs. execution. *Experimental Brain Research* **111**: 246–252.

Stich S and Nichols S (1992) Folk psychology: simulation or tacit theory? *Mind and Language* **7**: 35–71.

Wellman H (1990) *The Child's Theory of Mind*. Cambridge, MA: Bradford/MIT Press.

Wimmer H and Perner J (1983) Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* **13**: 103–128.

Further Reading

Carruthers P and Smith P (eds) (1996) *Theories of Theories of Mind*. Cambridge, UK: Cambridge University Press.

Davies M and Stone T (eds) (1995a) *Folk Psychology: The Theory of Mind Debate*. Cambridge, MA: Blackwell.

Davies M and Stone T (eds) (1995b) *Mental Simulation: Evaluations and Applications*. Cambridge, MA: Blackwell.

Goldman AI (1995) Simulation and interpersonal utility. *Ethics* **105**: 709–726.

Gordon R and Barker J (1994) Autism and the 'theory of mind' debate. In: Graham G and Stephens G (eds) *Philosophical Psychopathology*, pp. 163–181. Cambridge, MA: Bradford.

Heal J (1998) Co-cognition and off-line simulation: two ways of understanding the simulation approach. *Mind and Language* **13**: 477–498.