Reading Through Mirror Neurons? MindReading Reconsidered

Ali Yousefi Heris

Graduate School of
Systemic Neurosciences
LMU Munich

Munchen 2017

Reading Through Mirror Neurons? MindReading Reconsidered

Ali Yousefi Heris

Dissertation at the Graduate School of Systemic Neurosciences
Ludwig-Maximilians-Universitt München

Submitted by
Ali Yousefi Heris
Research Centre for Neurophilosophy and Ethics of Neuroscience

M̈unchen 2017

Supervisor: Prof. Dr. Stephan Sellmaier
Second advisor: Prof. Dr.-Ing. Stefan Glasauer
Third advisor: Prof. Stephen Stich
External Examiner: Prof. Shaun Gallagher
Date of Submission: 9 June 2017
Date of Defense: 23 September 2017

Acknowledgements

I wish to extend my sincere gratitude to:

My thesis advisors Stephan Sellmaier, Stefan Glasauer, and Stephen Stich for their constructive comments. Special thanks are due to Stephan Sellmaier for his guidance and continuous supports and encouragements throughout the project.

Thanks to the PhD students at the Research Centre for Neurophilosophy for sharing their thinking in our colloquiums.

Finally, my deepest gratitude goes to my family for their unflagging love and support.

Contents

0.1 Introduction

A central problem in cognitive sciences concerns the ability to represent our own or others' mental representations, what has been variously labeled as mindreading, theory of mind, folk psychology, or mentalizing. Many of our mental states represent how the world is, for example, the belief that "it is raining" represents a state in the world. Mental representations that are directed towards the world are called first-order representations. First-order representations, however, are themselves potential objects of representation. For example, "John knows that Mary loves him" is a second-order representation, and "John knows that Mary knows that he knows that Mary loves him" involves several levels of second-order representation. Mindreading, or mental state attribution, is a species of second-order representation that enables humans to explain, predict, and interpret behavior in mental terms. But how, exactly, this ability is acquired, and what kind of mechanisms underpin its operation? I will clarify these questions shortly, but let us first take a quick look at the philosophical background of the problem.

0.1.1 Other Minds and Mental State Terms

According to an influential view in philosophy of science, the criterion must be used to test the genuineness of scientific statements is the criterion of empirical verifiability. On this view, developed by logical positivists, scientific statements are factually significant if and only if the statements are, at least in principle, empirically verifiable. In Schlicks well-known slogan (Schlick, 1936), the meaning of an expression is the method of its verification. But if so, then how does the use of unobservable, theoretical entities in science square with this empiricist view? If empiricism is right about the science—that is if the source of the meaning of scientific statements lies in their relation to observation and experience—then what constitutes the semantics of theoretical terms such as "gravity", "electron", and "gene"? The same question arises in psychology. Mental states such as belief, desire, intention, pain and other psychological terms are theoretical, unobservable entities. What fixes the semantics of mental state terms in psychology?

One solution to this problem was behaviorism: the verification conditions for the meaning of mental state terms are behavioral. According to behaviorism, mental states do not refer to psychological episodes inside a person, rather, to say that a person is in a particular mental state simply means the same as saying, in an open-end list of statements, that the person is either behaving or has the disposition to behave in certain ways. The behaviorist position faced serious problems, perhaps the most notorious one was that the

account is circular. It was demonstrated that analyzing attribution of mental states in behavioral terms cannot be achieved unless our analysis makes reference to other mental states. The behaviorist account was replaced by a different view, developed by Lewis (Lewis, 1970, 1972), according to which theoretical terms get their meaning by being embedded in the theory in which they are used. On this account, theoretical terms have *functional* definition, that is, they are defined as the occupants of the causal roles specified by the theories within which they occur (Lewis, 1972, p. 254). So, the terms "gravity" and "electron" get their meaning by the causal roles the terms play in laws and generalizations of physics. Likewise, the semantics of mental states is constituted by the roles the terms play in the psychological theory in which they figure, a theory which has been invented long before the emergence of modern psychology. But what constitutes this folk psychological theory? Lewis remarks:

> Collect all the platitudes you can think of regarding the causal relations of mental states, sensory stimuli, and motor responses. ... Include only platitudes which are common knowledge among useveryone knows them, everyone knows that everyone else knows them, and so on ... and I am going to claim that names of mental states derive their meaning from these platitudes. (Lewis, 1972, p. 256)

So the content of the theory is constituted by a set of law-like generalizations that specify causal relations between mental states, sensory stimuli, and motor responses. The theory not only fixes the meaning of mental state terms and explains the regularities between sensory stimuli and behavioral responses, but also provide a solution to the other minds problem.

If, as Descartes believed, mind and body are fundamentally different (one located in space-time, the other not spatially located) and I am the only person who experiences my psychological states, then my belief in the existence of other minds might be seriously mistaken. Of course, we hardly doubt that others have a mental life, but what justifies this certainty? This has been one of the main and long-standing problems in the philosophy of mind. An immediate answer to this question is that we come to know what is going on in the others' mind by observing their behavior. However, the evidence (observed behavior) could have been produced by mindless bodies. It could be that others behave as if they are mindful, without actually experiencing any inner life. If so, then attribution of mental states to others is not easily *justified*.

How folk psychology solves the other minds problem? The convictions that others experience a mental life is not inferred from observing others'

behavior, nor it is derived by analogy from our access to our own mental life, but it can be seen as an *explanatory hypothesis* that, in conjunction with generalizations of folk psychology, provides explanation and prediction of others' behavior (Churchland, 1981, p. 69). Now, while mindreading is understood as a descendant of the other minds problem (Goldman, 2006; Nichols & Stich, 2003; Stich & Nichols, 2002), for researchers working on mindreading today there is nothing problematic about other minds. Rather, mindreading is seen as an "ability" that its execution serves the function of explaining and prediction others' behavior. The key question is that how mindreading is accomplished and what mechanisms underpin its operation.

0.1.2 Mindreading and Psychology

Empirical research on mindreading was launched by two primatologists, Premack & Woodruff (1978), who reported a series of problem-solving experiments on a chimpanzee named Sarah. In one experiment, videotapes of a human actor were presented who desired to eat bananas that were horizontally or vertically out of reach. Several photographs were presented, one of them was a solution to the problem, e.g. a stick to reach the bananas. Sarah solved the problem by consistently selecting the correct photographs, but the authors remarked that the subject's excellent performance was possible only if she could take into account the actor's mental states, specifically, the actor's intention and desire to reach the bananas. The study by Premack & Woodruff was followed by commentaries from a number of philosophers, including Bennett (1978), Dennett (1978), Pylyshyn (1978), and Harman (1978), remarking that we could never be sure that Sarah can think about another mind as long as her own mind is sufficient to solve the problem. What is required is to see if Sarah (or any creature for that matter) can make judgments about a target who has mental states different from her own, or by taking into account another's *false* belief. The commentaries resulted in devising a procedure that is now well known as the false-belief test, the acid test for the theory of mind.

0.1.2.1 The False-belief Test

The criterion suggested by philosophers was put into practice in a study by two psychologists, Wimmer & Perner (1983), who presented children a puppet show about two characters, Sally and Anne. Sally places his toy in a basket and leaves the scene. In his absence, Anne transfers the toy from the basket to a box. Children were then asked: when Sally returns, where will she look for his toy? The results show that children younger than 3

and a half fail to pass the test. Different versions of the test have been used in numerous studies, but they are all methodological variants of two basic procedures labeled as "unexpected content" or "unexpected transfer". The unexpected transfer is a variant of the original study by Wimmer & Perner (1983). In the unexpected content scenario, the child is shown a familiar container, often a tube of Smarties, and is asked what is inside. Then, contrary to the child's expectation, it is revealed that the real content is something quite different, pencils for example. Next, the child faces the test question: what someone else, the child who is outside the room and has not seen inside the box, will think is in the box. No matter which procedure is used, to give a correct answer in a false-belief test, the child must be able to set aside his own representations of reality and think about what the target thinks about a given situation, for example, to think what Sally thinks about the toy's location, and realize that Sally's action relies on a false belief, or misrepresentation, of reality. But how is this achieved? Theory-theory (TT) and simulation theory (ST) are the two dominant accounts developed to answer this question.

0.1.3 The Theory-Theory

According to one still dominant view, theory-theory, mindreading is guided and executed by a theory. In their commentaries on the chimpanzee's success, Premack & Woodruff (1978) argued that

> a system of inferences of this kind is properly viewed as a theory, first, because such states are not directly observable, and second, because the system can be used to make predictions, specifically about the behavior of other organisms. (Premack & Woodruff, 1978, p. 515)

Similarly, to account for the failure in the false belief test, it was argued that young children's failure and autistic subjects' difficulty with the test is the result of a not matured theory of mind (Leslie, 1987; Leslie & Frith, 1988). To succeed in the test, the child must be able to entertain a counterfactual situation, that "Sally believes that the toy is in the basket" even though she knows that the toy is not, in fact, in the basket. To achieve this, however, the child must possess a body of psychological knowledge, together with some processing mechanisms, such as the inference mechanism, that will put the psychological knowledge into use (Davies & Stone, 1995).

On TT account, we understand others "because we share a tacit command of an integrated body of lore concerning the law-like relations holding among external circumstances, internal states, and overt behavior"(Churchland, 1981,

p. 69). Moreover, the structural features of the theory "parallels perfectly those of mathematical physics; the only difference lies in the respect domain of abstract entities they exploit" (Churchland, 1981, p. 71). The laws represent various relations holding in the domain of psychology, which is adequate to the demands of everyday life. Everybody knows, for example, that "people who are angry are generally impatient", or if S desires P and believes that Q is a means to P, and S has no overriding desires, S will generally try to bring it about that Q (Churchland, 1981; Churchland & Churchland, 1998). However, it is notable that TT theorists disagree over the acquisition problem, some argue for an innate and modular version of TT (Leslie, 1994, 2000) whereas others argue that the theory is acquired in much the same way that scientific theories are acquired (Gopnik & Meltzoff, 1997; Perner, 1991).

Understanding others in terms of law-like generalizations seems problematic in at least two respects. Phenomenologically speaking, it does not seem that people understand and interact with others by going through an inferential process in which they deduce an explanandum (a mental state or a future behavior) by using a set of generalizations. In addition, frequent use of deductive processes seems cognitively too demanding. To avoid these problems, TT theorists have recently revised their view, characterizing the theory as an internally represented "knowledge structure" (Stich & Nichols, 1992) or model theories (Godfrey-Smith, 2005; Maibom, 2007, 2009), rather than law-like generalizations (Churchland, 1981). Understood this way, the processes operating to produce explanations or predictions do not need to be inferential, or at least nomological-deductive. In this more general sense, TT is understood as an information-rich process.

The theory-theory approach is a product of functionalism and has been the dominant explanatory strategy in understanding mindreading. (Stich & Nichols, 1992). In the late 1980s, however, a number of theorists laid down serious challenges to TT and developed an alternative view under the label of "simulation theory".

0.1.4 The Challenge from Simulation Theory

In the late 1980s, Gordon (1986) and, independently, Heal (1986) put forward the simulation hypothesis as an alternative to TT and functionalism. The view was further developed and defended by other theorists and psychologists, most notably Goldman (1989) and Harris (1992). Later, Goldman (2006) presented the most thoroughly developed and empirically informed defense of the simulation theory (ST). According to the advocates of the simulation hypothesis, understanding others is achieved by simulation, that is,

by imaginatively putting ourselves in the target's shoes, answering the question "what would I do in that person's situation?"(Gordon, 1986). Although details about how the simulation works slightly differ among simulation theorists, they all agree that we understand others by simulation, and without collapsing into theorizing, as long as two conditions are satisfied. First, the initial states of the simulating system are relatively similar to the mental states of the target. Second, the mechanism driving simulation is similar to the mechanism that drives the target (Goldman, 1989, 2006).

On a typical simulational account of action prediction, we put ourselves in the other's shoes and imagine what our mental states would be if we were in that target's situation. This "imaginative identification" (Gordon, 1992) or "transformation" (Gordon, 1996) would generate mental states that stand as representational surrogates for those of the target. The generated states, "tagged" as belonging to the target, are then fed into the decision-making or practical reasoning, mechanism. Because during simulation the decision-making system operates off-line—that is, it is disengaged from the motor control system and its normal operation—the output of the system, instead of generating action, is attributed to the target. Unlike TT, simulation is understood as an information-poor process.

The simulational account of mindreading has significant philosophical implications, in particular, it raises serious problems for functionalism and eliminativism. If mindreading is simulational and understanding others does not depend on a theory of mind, then the functionalist account of the meaning of mental state terms must be mistaken, because the assumption of a folk psychological theory that fixes the semantics of mental state terms would be explanatorily redundant. And, if there is no folk theory, it hardly makes sense to claim, as eliminativists do, that our common sense psychological theory is a defective theory that part or all of it will vanish in future.

In the last two decades, the debate between TT and ST has continued with highly sophisticated arguments on both sides; however, the progress was limited to articulating hybrid theories. However, by the turn of the century and following the discovery of mirror neurons there was a resurgence of interest in mindreading debates, in particular a revival of interest in the simulation hypothesis.

0.1.5 Mirror Neurons: Evidence from Cognitive Neuroscience

Two things happened by the turn of the century. First, whereas discussions in the early stages were mainly focused on propositional attitudes (beliefs, desires, intentions, etc.), a great deal of recent work on mindreading is devoted to the attribution of emotions and sensations. Second, positions

in those early days were mainly supported by evidence from developmental psychology, whereas recent theories, in particular, the simulational accounts, strongly rely on evidence from cognitive neuroscience. Specifically, the discovery of mirror neurons and mirror processes has opened a new angle into the mindreading debates.

One of the distinguishing features of (region) F5 neurons in the premotor cortex is that they code execution of goal-directed motor acts (Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992; Rizzolatti, Fogassi, & Gallese, 2001). However, it was incidentally discovered that a large proportion of F5 neurons fire in response to the presentation of motor acts, for example when the monkey only observes hand or mouth grasping (Pellegrino et al., 1992). Subsequent studies revealed that neurons in this region discharge both during observation and execution of the same action (Rizzolatti et al. 1996a; Gallese et al., 1996). Neurons with this property are dubbed mirror neurons because the brain in the observation mode behaves as if it mirrors the brain during the execution mode.

The discovery of mirror neurons has inspired a resurgence of interest in the simulation hypothesis. The link between mirror neurons and simulation was first created in a paper by Gallese (one of the neuroscientists who discovered mirror neurons) and Goldman published in 1998. Gallese and Goldman argued that, whereas the endogenous activation of F5 neurons is interpreted as constituting a "plan" to execute a certain action, the exogenous activation of the neurons, when observing the same action performed by other individuals, constitutes a "plan" which, instead of leading to execution, is attributed to the target. This is the way, they argued, we understand motor intentions. But why is the process simulational? Because the activity of mirror neurons "creates in the observer a state that matches that of the target. This is how it resembles the simulation heuristic." (Gallese & Goldman, 1998, p. 498). Besides, Gallese and Goldman argue that the evidence does not mesh with the TT account of mindreading because "nothing about TT leads us to expect this kind of matching" we see in mirror neurons (Gallese & Goldman, 1998, p. 498). Thus, the evidence from mirror neurons presents a basis for empirically discriminating between TT and ST.

The simulational interpretation of mirror neurons is advocated by other theorists as well. Fogassi et al. argue that mirror neurons not only code the observed motor act but also allow the observer to understand the agent's intentions (Fogassi et al., 2005, p. 662), and Iacoboni et al. (2005) consider mirror neurons as mechanisms responsible for understanding intentions. The claim, however, is not limited to motor intentions. Mirror neurons are discovered across different domains, including sensations and emotions of fear (Adolphs et al., 2005), anger (Lawrence, Calder, McGowan, & Grasby,

2002), pain (Singer et al., 2004), and touch (Keysers et al., 2004). Based on this evidence, a number of theorists have argued that emotion and sensation recognition is also simulational. Gallese et al. maintain that mirror mechanisms "allow us to directly understand the meaning of the actions and emotions of others by internally replicating (simulat-simulating) them without any explicit reflective mediation" (Gallese, Keysers, & Rizzolatti, 2004, p. 396). Among the theorists, Goldman (2006) has provided the most comprehensive account of simulation, both at low-level, mirror-based recognition of emotions and sensations and at high-level, imagination-based, attribution of propositional states.

While the discovery of mirror neurons has rejuvenated the debates and attracted a significant group of simulation advocates, several problems in the literature remain unaddressed, including questions concerning the concept of simulation, the precise role of mirror neurons in mindreading, the way mirror neurons support ST, the reasons the processes do not square with TT, and examination of the explanatory value of TT and ST in domains where mindreading is exhibited. These questions are of prime interest in my agenda.

0.1.6 The Central Hypotheses

This dissertation consists of five stand-alone sections, each dealing with a different hypothesis on mindreading. Beginning with interpretation questions, section one examines the simulation hypothesis at the conceptual level. How the ST theory, and the notion of simulation on which the theory relies, ought to be understood. After evaluating several candidates, I argue that the most promising senses of simulation, the similarity-based and the off-line simulation, fail to discriminate simulational from non-simulational processes.

Simulation is a process, and a process P simulates another process P′ only if P duplicates, or resembles P′ in some significant respects. But what are the likely dimensions of resemblance between P and P′? Likely respects of resemblance have already been discussed in the literature—process similarity, concrete similarity, phenomenological, and functional similarity—all of them, however, face serious problems (Section One). More recently, however, ST theorists have argued that neurological similarity is the most promising respect of resemblance in simulation. In section two, I criticize this view and show that the processes involved in a class of celebrated simulation prototypes do not show neurological similarity in the sense simulation theorists contend.

Drawing on evidence from brain imaging and lesion studies, simulation theorists have argued that in recognizing an emotion we use the same neural

processes used in experiencing that emotion. In section three, I argue that the view is fundamentally misguided. To show this, I will examine the simulational arguments for the three basic emotions of fear, disgust, and anger, and argue that the simulational account relies strongly on a narrow sense of emotion processing which hardly squares with evidence on how, in fact, emotion recognition is processed. I contend that the current body of empirical evidence suggests that emotion recognition is processed in an integrative system involving multiple cross-regional interactions in the brain, a view which also squares with understanding emotion recognition as an information-rich, rather than simulational, process.

Section four examines the simulation hypothesis in connection with communication, in particular, examine the explanatory power of ST in accounting for the mindreading exhibited in utterance interpretation. I discuss several problems with the simulation hypothesis, most importantly I argue that the simulation strategy is not only cognitively too demanding but virtually ineffective in communicative contexts. Moreover, drawing on empirical evidence from three clinical populations, I show that deficits in pragmatic interpretation are not associated with simulation impairments. Therefore, I argue that simulation cannot play any significant role in utterance interpretation.

Section five links mindreading to moral psychology. For centuries, it is believed that an agents action shows weakness of will if he acts, freely and intentionally, counter to his own assessment of the action. In this final section, I provide a framework for a more natural and empirically oriented account of weakness of will. Relying on evidence from developmental psychology, I argue that weakness of will is essentially nothing but an exercise in metarepresentation and intention recognition.

0.1.7 Bibliography

Adolphs, R., Gosselin, F., Buchanan, T. W., Tranel, D., Schyns, P., & Damasio, A. R. (2005). A mechanism for impaired fear recognition after amygdala damage. Nature, 433(7021), 68-72.

Bennett, J. (1978). Some remarks about concepts. Behavioral and Brain Sciences, 1(4), 557-560.

Churchland, P. M. (1981). Eliminative Materialism and the Propositional Attitudes. The Journal of Philosophy, 78(2), 67-90.

Churchland, P. M., & Churchland, P. S. (1998). On the Contrary: Critical Essays, 1987-1997. Cambridge: MIT Press.

Davies, M., & Stone, T. (1995). Folk psychology: the theory of mind debate. Oxford; Cambridge, Mass.: Blackwell.

Dennett, D. C. (1978). Beliefs about beliefs [P&W, SR&B]. Behavioral and Brain Sciences, 1(4), 568-570.

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., & Rizzolatti, G. (2005). Parietal Lobe: From Action Organization to Intention Understanding. Science, 308(5722), 662-667.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. Brain, 119(2), 593-609.

Gallese, V., & Goldman, A. I. (1998). Mirror neurons and the simulation theory of mind-reading. Trends in Cognitive Sciences, 2(12), 493-501.

Gallese, V., Keysers, C., & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. Trends in Cognitive Sciences, 8(9), 396-403.

Godfrey-Smith, P. (2005). Folk psychology as a model. Philosophers Imprint, 5(6), 1-16.

Goldman, A. I. (1989). Interpretation Psychologized. Mind & Language, 4(3), 161-185.

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

& Meltzoff, A. N. (1997). Words, Thoughts, and Theories. Cambridge, Mass.: A Bradford Book.

Gordon, R. M. (1986). Folk Psychology as Simulation. Mind & Language, 1(2), 158-171.

Gordon, R. M. (1992). Reply to Stich and Nichols. Mind & Language,

7(1–2), 87-97.

Gordon, R. M. (1996). Radical Simulationism. In P. Carruthers & P. K. Smith (Eds.), Theories of Theories of Mind. Cambridge University Press.

Harman, G. (1978). Studying the chimpanzees theory of mind. Behavioral and Brain Sciences, 1(4), 576-577.

Harris, P. L. (1992). From Simulation to Folk Psychology: The Case for Development. Mind & Language, 7(1-2), 120-144.

Heal, J. (1986). Replication and Functionalism. In J. Butterfield (Ed.), Language, Mind, and Logic (Vol. 14, pp. 135150). Cambridge University Press.

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., & Rizzolatti, G. (2005). Grasping the intentions of others with ones own mirror neuron system. PLoS Biology, 3(3), e79.

Keysers, C., Wicker, B., Gazzola, V., Anton, J.-L., Fogassi, L., & Gallese, V. (2004). A Touching Sight: SII/PV Activation during the Observation and Experience of Touch. Neuron, 42(2), 335-346.

Lawrence, A. D., Calder, A. J., McGowan, S. W., & Grasby, P. M. (2002). Selective disruption of the recognition of facial expressions of anger. Neuroreport, 13(6), 881-884.

Leslie, A. M. (1987). Pretense and representation: The origins of theory of mind. Psychological Review, 94(4), 412-426.

Leslie, A. M. (1994). ToMM, ToBy, and agency: Core architecture and domain specificity. In L. A. Hirschfeld & S. A. Gelman (Eds.), Mapping the Mind: Domain Specificity in Cognition and Culture (pp. 119-148). Cambridge University Press.

Leslie, A. M. (2000). Theory of mind as a mechanism of selective attention. In M. Gazzaniga (Ed.), The New Cognitive Neurosciences (pp. 1235-1247). Cambridge, MA, MIT Press.

Leslie, A. M., & Frith, U. (1988). Autistic childrens understanding of seeing, knowing and believing. British Journal of Developmental Psychology, 6(4), 315-324.

Lewis, D. (1970). How to Define Theoretical Terms. Journal of Philosophy, 67(13), 427-446.

Lewis, D. (1972). Psychophysical and Theoretical Identifications. Australasian Journal of Philosophy, 50(3), 249-258.

Maibom, H. (2007). Social Systems. Philosophical Psychology, 20(5), 557-578.

Maibom, H. (2009). In Defence of (Model) Theory Theory. Journal of Consciousness Studies, 16(6-1), 360-378.

Nichols, S., & Stich, S. P. (2003). Mindreading. An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds. Oxford University Press.

Pellegrino, G. di, Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. Experimental Brain Research, 91(1), 176-180.

Perner, J. (1991). Understanding the representational mind (Vol. xiv). Cambridge, MA, US: The MIT Press.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences, 1(4), 515-526.

Pylyshyn, Z. W. (1978). When is attribution of beliefs justified? [P&W]. Behavioral and Brain Sciences, 1(4), 592-593.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. Cognitive Brain Research, 3(2), 131-141.

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. Nature Reviews Neuroscience, 2(9), 661-670.

Schlick, M. (1936). Meaning and Verification. Philosophical Review, 45(4), 339-369.

Singer, T., Seymour, B., ODoherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for Pain Involves the Affective but not Sensory Components of Pain. Science, 303(5661), 1157-1162.

Spaulding, S. (2012). Mirror neurons are not evidence for the Simulation Theory. Synthese, 189(3), 515-534.

Stich, S. P., & Nichols, S. (1992). Folk Psychology: Simulation or Tacit Theory? Mind & Language, 7(1–2), 35-71.

Stich, S. P., & Nichols, S. (2002). Folk Psychology. In S. P. Stich & T. A. Warfield (Eds.), Encyclopedia of Cognitive Science (pp. 35-71). Blackwell.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and

constraining function of wrong beliefs in young childrens understanding of deception. Cognition, 13(1), 103-128.

0.2 The Concept of Simulation

Abstract

Simulation theorists argue that we can use our own mind as a model to understand other minds. In spite of the attractiveness of the hypothesis, it remains fairly unclear how the theory, and the notion of simulation on which it depends, ought to be understood. I discuss different senses of simulation and argue that the most promising sense of simulation, the similarity-based and the off-line senses of simulation, are faced with several problems; most importantly, they fail to discriminate simulation from non-simulational processes.

0.2.1 Respects of Similarity

How is mental state attribution executed? According to one still dominant view, the *theory-theory*, mental state attribution, or mindreading, is underpinned by a set of mental representations and processes that operate on these representations. Mindreading, on the theory-theory account, is understood as an information-rich process. However, in the last two decades or so the theory-theory position is challenged by a rather different view, the *simulation theory*, whose advocate hold that mindreading is subserved by a process of simulation in which, instead of relying on a theory of target, we use our own mind as a model for the other person's mental life. Mindreading on the simulational account is basically understood as an information-poor process (Davies & Stone, 1995a, 1995b; Nichol & Stich, 2003).

Although the simulation theory is supported by a considerable number of theorists, even a cursory glance at the literature reveals that the notion of simulation is understood in heterogeneous ways such that so many of the processes that the advocates of simulation theory see as simulational have no or hardly anything in common. For instance, Robert Gordon, one of the original advocates of the simulation theory, understood simulation in terms of "imaginative identification" (Gordon, 1992) and "imaginative transformation" (Gordon, 1996). Paul Harris (1992) and Jane Heal (1996) supported a version of simulation which is not imagination-based, but rather involves a process in which a mindreader simulates by actually placing himself in a situation which is very similar to the target's. And, Goldman (Goldman, 1989, 2006) proposed a notion of simulation which, although similar to Gordon's is imagination-based, operates in a particular way in which the simulating system is taken off-line and is provided with pretend inputs, the so-called off-line simulation.

So, despite the theory's wide range of applications and its attractiveness to a significant group of advocates, it remains fairly unclear how the theory,

and the notion of simulation on which it relies, ought to be understood. In reaction to this situation, Stich and Nichols (Nichols & Stich, 2003; Stich, 2009; Stich & Nichols, 1997) have frequently remarked that the diversity among the processes to which simulation theorists have attached the label simulation "is so great that the term itself has become quite useless. It picks out no natural or theoretically interesting category" (Stich & Nichols, 1997, p. 299). However, Goldman, the most prominent defender of the simulation theory, resists this complaint and insists that "there is unity amid this diversity; simulation is still a natural and theoretically interesting category. Analogously, although there are many different atomic elements, the category "atomic element" is a natural and theoretically interesting category." (Goldman, 2006, p. 35).

Since Goldman "does not want to defend every application of the term 'simulation' that anybody has ever proposed" (Goldman, 2009c, p. 138), he sets out his own definition of simulation. So, if P and P′ are two processes, what is required for P to qualify as a simulation of P′? Here is Goldman's first pass at defining generic simulation:

> Generic Simulation (initial): Process P is a simulation of another process P′ = df.
> P duplicates, replicates, or resembles P′ in some significant respects (significant relative to the purposes of the task). (Goldman, 2006, p. 36)

As Goldman remarks, P and P′ are token processes rather than process types, and P can have any temporal relation to P′. Further, the simulated activity P′ might be only hypothetical rather than actual—as when a flight simulator simulates a crash that never corresponds to a real crash. One problem, Goldman notes, with this initial definition of generic simulation is that "duplication, or resemblance, is symmetrical, whereas simulation is not" (Goldman, 2006, p. 37). An actual flight might resemble or duplicate a flight simulation, but it doesn't simulate what happens in a flight simulator. This suggests that, on Goldman's view, there is more to simulation than mere duplication. In the case of mindreading, he remarks, it is the mental activity of the mindreader (simulator) that simulates that of the target, not the other way around. One way to get around this problem would be as follows:

> Simply require that the simulating process occur out of the *purpose*, or *intention*, to replicate the simulated process ... This won't quite work, however, because it is doubtful that all simulation is purposeful. Some simulation may be automatic and nonpurposeful. (Goldman, 2006, p. 37)

To overcome the problem, Goldman suggests that, even without purposefulness, one phenomenon count as a simulation of another "if it is the *function* of the former to duplicate or resemble the other" (Goldman, 2006, p. 37). This revises the initial definition of generic simulation as follows:

> Generic Simulation (revised): Process P simulates process P′ = df.
> (1) P duplicates, replicates, or resembles P′ in some significant respects (significant relative to the purposes or function of the task), and
> (2) in its (significant) duplication of P′, P fulfills one of its purposes or functions. (Goldman, 2006, p. 37)

Generic simulation applies to both mental and nonmental processes. So, Goldman defines mental simulation as follows:

> Mental Simulation: Process P is a mental simulation of target process P′ = df.
> Both P and P′ are mental processes (though P′ might be merely hypothetical), and P and P′ exemplify the relation of generic simulation as previously defined. (Goldman, 2006, p. 37-8).

This definition, although developed with care and in minute detail, is still rather vague and offers no real help. It is fairly unclear what is precisely intended by "duplication" or "resemblance" in (1), and what the terms "purpose" and "function" exactly designate in (2). So lets clarify (1) and (2) each in turn.

Let P and P′ be fundamentally different systems. P might be an abstract entity such as a scientific model (e.g. the Bohr model of the atom) in which an identifying description allows for a surrogative reasoning about the target. But some abstract models need to be physically implemented and experimented upon in order to fulfill their representational function. For example, computer simulations are of mathematical nature, but are realized and run as a program in a computer in order to solve equations that represent the time-evolution of a target system. What kind of resemblance is involved here? At the physical level of description, a computer simulation bears no physical resemblance to its target. There is no physical resemblance between the computer simulations of, for example, human brain and particular human brains. Furthermore, computer simulations and simulated system do not work according to the same rules and principles (Haugeland, 1989). The rules that govern simulation of human brain differ from the rules that govern operations of particular brains. Given the dissimilarities, how does a

computer simulation represent? It neither goes through the states that transpires in the target nor follows the same principles that govern operations in the target. Rather, the success of simulation in this case relies on a theory or an accurate description of the target. As Goldman puts it, "if a computer or a person seeks to simulate a system fundamentally different from itself (e.g., a weather system or an economy), it must be driven by a good theory of that target. Let us call this theory-driven simulation." (Goldman, 2006, p. 32). So, simulation would be theory-driven if P and P′ stand for processes in fundamentally different systems. But theory-driven duplication is not the intended sense of resemblance in (1).

To avoid theory-driven simulation, let P and P′ be similar, rather than different, systems. For instance, there are cases, where either there is no precise theory of how a target system works or if there is, it is either impractical or too complicated to use the theory. In such cases, a rather different strategy is to use a second system that stands for the target system. For example, scientists very often use monkey brains as a model organism to learn more about the human brain. In this case, the monkey brain represents the human brain by following the same principles and undergoing the same states that occur in the human brain. In other words, as far as the success of the modeling is concerned, the monkey brain represents the human brain by concrete resemblance (Fisher, 2006), rather than by appealing to a theory of the target. This form of process-driven (Goldman, 1989) simulation might be the relevant respect of similarity in (1). Thus, (1) can be slightly modified as P simulates P′ only if P concretely duplicates or resembles P′.

When two systems show concrete similarity, it seems that the simulating system is like a faithful replication of the target. But there is no such thing as a perfectly faithful replication; replication, even in concretely similar systems, is always restricted to some respects. To put it differently, a simulating system might resemble the target in some respects and still be different in others. This raises the question of which respects of resemblance are significant for the purpose of simulation. One very outstanding respect of similarity in simulation strategy is phenomenology. Here is how phenomenological respect figures in simulationists' arguments:

> The simulation idea has obvious initial attractions. Introspectively, it seems as if we often try to predict others' behavior—or predict their (mental) choices—by imagining ourselves in their shoes and determining what we would choose to do. (Goldman, 1989, p. 169)

To see the point, consider, for example, a study by Kahneman and Tversky (1981), where subjects were asked to consider two travelers, Mr. Crane

and Mr. Tees, who were scheduled to leave the airport on different flights, at the same time. They traveled from town in the same limousine, which was caught in a traffic jam, and arrived at the airport 30 minutes late. Mr. Crane is told his flight left on time. Mr. Tees is told that his flight was delayed, and just left 5 minute ago. Who do you think is more upset? 96% of subjects said that Mr. Tees was more upset. How did they come up with this answer? Goldman remarks that, "each subject would have put himself in each of the imaginary traveler's shoes and imagined how he would have *felt* [emphasis added] in that place" (Gallese & Goldman, 1998, p. 496). So, in understanding the target, each subject simulates by undergoing an emotional contagion that bears phenomenological resemblance to what happens in the target. Similarly, Gordon (1986) maintains that in predicting the behavior of others, we simulate by answering the question 'what would I do in that person's situation?'. This involves, on Gordon's account, to imaginatively project into the other's situation in the same way that chess players in playing against an opponent, while 'transported in imagination', visualize the board form the other side and act accordingly. Thus, the idea of putting oneself in the other's shoes involves mindreader to be engaged in a form of 'empathetic understanding'(Gordon, 1986) and forming mental imageries which are associated with phenomenological resemblance to what happens in the target.

Phenomenological resemblance, despite its initial attraction, confronts serious problems. First, although studies show phenomenological resemblance during, for example, introspective imagery (Kahneman & Tversky, 1981), or observation and attribution of some emotional facial expressions (Wicker et al., 2003), the cases are restricted only to the attribution of perceptual, emotional or sensational states. But mindreading often involves attribution of other types of mental states including beliefs, intentions, desires and all the so-called propositional states. Propositional states, however, do not have perceptual or emotional content, thus hard to see how their attribution might be associated with phenomenological properties. Second, phenomenology is absent even in the attribution of emotional and sensational states. Recently, Goldman (2006) has distinguished between high-level and low-level mindreading, and argued that low-level mindreading is the automatic and unconscious process of detecting emotional and sensational states. How can something that happens automatically and subconsciously be associated with phenomenological properties? There is indeed a general consensus among theorists that mindreading largely occurs below the threshold of consciousness. If so, then phenomenology is not always associated even with the attribution of emotional and sensational states. What about cases like the above studies by Kahneman and Tversky (1981) or Wicker et al.

(2003) that simulation theorists cite? Well, we must ask: is phenomenology conceived as not compatible or incapable of harmonious combination or co-existence with other non-simulational processes? If not, then even in cases where mental state attribution is associated with phenomenological resemblance, the resemblance in question does not necessarily entail simulation. That is, despite showing phenomenological resemblance, it might turn out that the attribution in those cases is guided and executed by a theory-driven or other type of non-simulational process. I conclude that phenomenology cannot be the intended respect of similarity in Goldman's definition in (1).

What might be the possibly more relevant respects of resemblance in simulation? Goldman notes, "the respects of resemblance I shall highlight are functional or neural" (Goldman, 2006, p. 151). Beginning with functional, or input/output, similarity, several problems arises immediately. To begin with, the simulation prototypes, for instance, the simulational accounts of action prediction (Goldman, 1989, 2006), predicting grammatical judgment (Harris, 1992) and inference prediction (Nichols & Stich, 2003; Stich & Nichols, 1995) are all accounts that rely on functional dissimilarities. For instance, on the standard simulational account of action prediction, we take our own decision-making system off-line and provide it with pretend inputs, and let it to output a decision which we subsequently attribute it to the target. The crucial point for our discussion is that the decision-making system during simulation, compared to its standard (non-mindreading) operation, is taken off-line and are provided with pretend inputs.[1] Since the inputs are pretend, that is, tagged as belonging to the target, they are associated with a pattern of causal connections—connections to other mental states, internal mechanisms, and output behavior—which is functionally different from the pattern associated with genuine, non-pretend, inputs. As a result, the mechanism that subserves simulation runs in a way different from the way it does in its standard mode of operation. So, whereas during the online operation the decision-making system takes standard inputs and returns a genuine decision, during simulation the system operates off-line and returns an output which is disengaged from mindreader's behavior. This shows why the simulating system and target don't resemble each other in functional terms. The same functional dissimilarities hold in Harriss simulational account of grammaticality prediction (1992), or Stich and Nichols' simulational account of inference prediction (Nichols & Stich, 2003; Stich & Nichols, 1995).

So functional similarity does not hold in the prototype accounts of simulation theory. But even if it does, it is not sufficient for a process to qualify as

[1]Not all simulation theorists are committed to the off-line or pretense-drive simulation. See, for instance, Harris (1992) and Heal (1996).

simulational. Simulation theorists hold that a process counts as simulational if the initial states and the simulating mechanism resemble to those of the target. Here is how Goldman makes the point:

> In the mindreading case, process-driven simulation can succeed in producing a final state that is identical or isomorphic to that of the target as long as (1) the process or mechanism driving the simulation is identical, or relevantly similar, to the process or mechanism that drives the target and (2) the initial states of the simulating system (the attributor) are the same as, or relevantly similar to, those of the target. Process-driven simulation does not collapse into theorizing. (Goldman, 2006, p. 32)

Similarely, Currie & Ravenscroft remark:

> It is a feature of simulative processes that the mechanism which underpins the simulative process is of the same *type* as the mechanism which is being simulated. Thus, according to the simulation account of the capacity to predict and explain action, the simulative process involves the predictor's decision-maker which is type-identical to the target's decision- maker: they are both tokens of the type 'normal human decision-maker'. (Currie & Ravenscroft, 1997, p. 164)

However, a simulating system might resemble the target by these definitions and nevertheless still not qualify as a simulational process (Ramsey, 2010). To see this, consider the case where Mr. A believes that all philosophers are shy. Mr. A is introduced to Mrs. B and he is told that she is a philosopher. You are asked to predict what Mr. A will say if asked whether Mrs. B is shy. How would you proceed to make this prediction? On one account, since this is an inference prediction, it begins by feeding the inference mechanism with pretend inputs—that is, the assumptions that you think the target holds, including the beliefs that 'all philosophers are shy' and 'Mrs. B is a philosopher'. The inference mechanism outputs that Mrs. B is shy. You attribute the output to the target and come to believe that Mr. A thinks that Mrs. B is shy. The process must count as simulational because, as stipulated above, the initial states and the mechanism that drives the simulation resemble to those of the target.[2]

However, this form of similarity holds even in a non-simulational or information-rich account of your prediction. On an information-rich account,

[2]This is a modified version of what Stich and Nichols (1995) raise in their discussion of type-2 Harris simulation.

you begin by entertaining the relevant assumptions including Mr. A's beliefs that 'all philosophers are shy', and 'Mrs. B is a philosopher'. Since this is an inference prediction, the assumptions are fed as input into the inference mechanism. The inference mechanism, provided with the inputs, along with the aid of a tacit theory of reasoning, returns that Mr. A thinks that Mrs. B is shy. Thus, input/output similarity is not even sufficient for simulation.

0.2.2 Neurological Similarity and Simulational Models

One way to avoid the above problem is to look for a stronger conception of similarity at a lower level of description that goes beyond input/output similarity. One serious possibility here is neurological resemblance. Several findings from cognitive neuroscience show an overlap in neuronal discharge between different modes of endogenous and exogenous activation during, for instance, motor actions (Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992; Rizzolatti, Fadiga, Gallese, & Fogassi, 1996), sensation and emotional responses in fear (Adolphs et al., 1999; Adolphs, Tranel, Damasio, & Damasio, 1994), disgust (Phillips et al., 1997; Wicker et al., 2003), pain (Jackson, Meltzoff, & Decety, 2005; Singer et al., 2004), and anger (Lawrence & Calder, 2004; Lawrence, Calder, McGowan, & Grasby, 2002), among others. In each case, the brain during the observation mode activates as if it is mirroring the activity of the brain during the execution/experience mode. For example, in the case of motor mirror neurons, observing a hand grasping an apple causes the observer's motor system to resonate, in the sense that the same population of neurons (in the premotor cortex) that controls the grasping of an apple also becomes active during the observation of the grasping movement. Likewise, there seems to be a neurological resemblance between the experience and recognition of certain emotions and sensations (more discussion on this below). This presumably can turn the above formulation into a more promising shape in which P simulates P′ only if P resembles, or duplicates, P′ in neurological terms. Neurological resemblance, however, faces three serious problems.

First, mental simulations are used to answer questions or form beliefs about other people (Goldman, 2006, p. 39). This presumably implies that simulation, whether it is used for mindreading or other purposes, has an epistemic purpose, that is, it facilitates acquiring information about a target system. However, consider the cases when, for instance, when sports fans who, while observing an athlete doing an exercise, tend to 'help'him by mimicking the athlete's movements (Rizzolatti, Fogassi, & Gallese, 2001). In such situations, whereas similar neural substrates are implicated during endogenous and exogenous activations, the mimicry performed by sports fans

hardly counts as processes that serving any epistemic purpose. If that is the case, then there are situations where we have neurological resemblance without having simulation. So, if P simulates P in terms of neurological resemblance, the problem arises as to how to discriminate processes which are simulational, and so have an epistemic purpose, from processes such as athletes' mimicry which show neurological resemblance, but have no epistemic purpose whatsoever.

Second, neural duplications, even when they serve a cognitive function, may be involved in a large variety of purposes. For example, imitation is the automatic tendency to reproduce an observed action. At the neurological level, it is suggested that imitation is subserved by mirror neurons that, on the one hand, code motor acts, and on the other, allow imitation to take place (Rizzolatti & Craighero, 2004; Rizzolatti et al., 2001). This suggests that imitation involves neurological resemblance. But it is notable that humans are capable of *learning* by imitation (Rizzolatti & Craighero, 2004). This again shows a case of neurological resemblance which is not a case of simulation. Advocates of the simulation theory would probably argue that simulation is not restricted to mindreading, but rather is involved in a variety of cognitive functions, including learning by imitation. This solution, however, doesn't seem to be helpful as in that case it would be hard to see how simulation theory can discriminate simulation for mindreading from its other functions, for example, simulation for the purpose of imitation.

Third, neurological similarity is not sufficient for simulation. In other words, a process might resemble, in neurological terms, a target process and nevertheless qualify as an information-rich rather than simulational process. Indeed, studies do suggest similarity, in neurological terms, between vision and visualization, or action and motor imagination.[3] However, in most cases, visualization or motor imagery is under subject's voluntary control and is often driven by information about the target's situation. For instance, subjects learn about the target by experimenters' description, and even might be instructed on how to build up a mental representation of the target (Decety, Jeannerod, & Prablanc, 1989). Sometimes the tasks involve making inferences (Schwartz & Black, 1999), or is such that it taps into subjects memory and knowledge (Meudell, 1971).[4] Theory-driven neurological duplications, however, do not qualify as simulational. So here again, there are cases where we have neurological resemblance without having simulation. The problem

[3]See, for example, Jeannerod (2001) for neural similarity in motor imagination, and Kosslyn et al. (1999) for visualization.

[4]The question as to whether visualization involves the same perceptual representations as vision is quite contentious and has been the subject of debate in the literature. See, for example, Pylyshyn (1973) and Farah (1988).

then arises as to how simulation theory discriminates neurological duplications which are simulational from those which are information-driven.

Should we drop the similarity-based characterization of simulation? Not any of the phenomenological or the functional forms of similarity seem to be the intended respect of resemblance in (1), and the neurological resemblance, as the most promising form of similarity, fails to discriminate simulation from non-simulational processes. One way to avoid these problems might be suggested by (2), where it has been remarked that P in its duplication of P′ fulfills one of its *functions*. So, as it has been noted, "a phenomenon intuitively counts as a simulation of another if it is the function of the former to duplicate or resemble the other" (Goldman, 2006, p.37). This suggests that P simulates P′ only of P resembles, or duplicates, P′ in neurological terms, and P in its neurological duplication performs its function. This suggestion, however, seems to be of little help. How are we to make sense of the notion of function here? What is it for P to have the function of (neurologically) duplicating, or resembling, P′? To be sure, the purpose or function of simulation is to achieve mindreading. The question, rather, is that how exactly a process serves this function, and what is required for a process, in addition to neurological resemblance, to carry out this function? We are not given a clue. Goldman remarks that "I lack a theory of function to provide backing for this approach, but I shall nonetheless avail myself of this notion" (Goldman, 2006, p. 37).

To get around this problem, it might be suggested that we should take into account that a distinguishing feature of simulation is that a mindreader makes special use of her own mind (or brain) in assigning mental states to others (Goldman, 2006, p.40). Specifically, it might be suggested that simulation involves that the simulating system operates off-line and is provided with pretend or non-standard inputs. So, P simulates P′ only if P, in addition to bearing neurological resemblance to P′, operates off-line and is provided with pretend, or non-standard, inputs.

This option, however, does not seem to be helpful, because simulational processes, at least for low-level mindreading where simulation theory can find evidence for neurological similarities, neither operate off-line nor are provided with pretend inputs. First, off-line operation at low-level mindreading involves the simulating process to be momentarily disengaged from the mindreader's emotion, sensation, or motor control systems. However, none of the four (low-levely mindreading) models proposed by simulation theorists posit any off-line operation during the simulation process: 1) the Generate and Test, 2) the Reverse Simulation, 3) the Simulation with as-if Loop, and 4) the Unmediated Resonance (Mirroring) (Goldman, 2006b; Goldman & Sripada, 2005). To see this, let us very briefly review the operation of the

four models.

In the Generate and Test model, generating a hypothesized emotion prompts its natural facial expression in the mindreader; in the Reverse Simulation model, visual representation of the target's facial expression gives rise to the activation of the mindreader's facial muscles which consequently generates an experience of the corresponding emotional state; in the Simulation with as-if Loop model, there is a link between a visual representation of a targets facial expression and a somatosensory representation of 'what it would feel like'to make that expression, which generates the experience of the corresponding emotion. Finally, in the Unmediated Resonance model, observation of the target's facial expression directly triggers activation of the same neural substrate of the emotion in question, and as a result experience of the corresponding emotion (Goldman, 2006b; Goldman & Sripada, 2005). So the simulational processes, even when there is a neurological resemblance, don't operate off-line. Mirror neuron activation and mirror processes, according to these models, is not disengaged but rather impacts the observers facial expressions and his emotional or sensational system. As a result, the mirroring process generates an output which is the genuine experience of an emotion, not a pretend output. This generation of a genuine output is emphasized by simulation theorists. For example, Wicker et al. (2003) remark that mirror neuron activation generates an emotional contagion, a feeling of disgust, which must occur in the observer in order to understand the facial expression of disgust. Similarly, Gallese (2001) suggests that understanding emotions and sensations of others requires a self-other identity relation which, in addition to the mirror activation, entails an emotional contagion.

I argued that mirror neurons don't operate off-line, but mirror neurons don't operate on pretend inputs, either. Pretend inputs differ from the genuine ones only in that they are 'tagged' as belonging to the target, but they resemble the genuine inputs in that the pretend inputs are representational surrogates for the inputs in the target (simulated) process. Indeed, this is a requirement otherwise simulation would lead to less accurate or mistaken results. However, this input commonality is absent in the endogenous and exogenous activation of mirror neurons and processes. For example, during action observation, the inputs into the parieto-frontal circuit arrive from higher order visual areas, such as the superior temporal sulcus, whereas during action execution, they mostly come from the temporal lobes (Rizzolatti & Sinigaglia, 2010, p. 265). To put it more intuitively, whereas mirror neurons during action observation are triggered by inputs from visual representation of action and exteroceptive information, action execution is prompted by proprioceptive and interoceptive inputs.

Moreover, a comparison between the endogenous and exogenous activa-

tion of motor mirror neurons suggests that the neurons are only broadly congruent in their responses to the observation and execution of effective actions (Gallese, Fadiga, Fogassi, & Rizzolatti, 1996). For example, two-thirds of the F5 mirror neurons, in order to be triggered, don't need the observation of the action that they code during action execution. More interesting are a set of mouth mirror neurons, called communicative neurons, which code execution of ingestive actions, but the most effective action for them during observation are communicative gestures (Rizzolatti & Craighero, 2004). So, how can a set of neurons which code for a certain action, but visually respond to a different action, have similar inputs under different modes of exogenous and endogenous operation?

Taken together, mirror neurons neither run off-line, nor operate on pretend inputs/outputs. Maybe, as it has been remarked (Goldman, 2009c, p. 149), generalizing off-line simulation to low-level mindreading is too narrow and constraining. In that case, it seems that simulation theory is caught in a thorny dilemma: either we keep off-line simulation within the bounds of high-level mindreading and define simulation in terms of neural resemblance only, in which case we face the above discussed problems; specifically, we fail to discriminate simulation from non-simulational processes. Or, to avoid those problems, we generalize off-line simulation to the low-level mindreading and add it to the neurological similarity condition, in which case mirror-base mindreading would not qualify as simulational because the off-line description would be too narrow and constraining.

0.2.3 Bibliography

Adolphs, R., Tranel, D., Damasio, H., & Damasio, A. (1994). Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. Nature, 372(6507), 669-672.

Adolphs, R., Tranel, D., Hamann, S., Young, A. W., Calder, A. J., Phelps, E. A., ... Damasio, A. R. (1999). Recognition of facial emotion in nine individuals with bilateral amygdala damage. Neuropsychologia, 37(10), 1111-1117.

Davies, M., & Stone, T. (1995a). Folk psychology: the theory of mind debate. Oxford; Cambridge, Mass.: Blackwell.

Davies, M., & Stone, T. (1995b). Mental Simulation: Evaluations and Applications - Reading in Mind and Language (1 edition). Oxford, UK; Cambridge, Mass: Wiley-Blackwell.

Decety, J., Jeannerod, M., & Prablanc, C. (1989). The timing of mentally represented actions. Behavioural Brain Research, 34(1-2), 35-42.

Farah, M. J. (1988). Is visual imagery really visual? Overlooked evidence from neuropsychology. Psychological Review, 95(3), 307-317.

Fisher, J. C. (2006). Does Simulation Theory Really Involve Simulation? Philosophical Psychology, 19(4), 417-432.

Gallese, V. (2001). The 'Shared Manifold" Hypothesis: From Mirror Neurons to Empathy. Journal of Consciousness Studies, 8(57), 33-50.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. Brain, 119(2), 593-609.

Goldman, A. I. (1989). Interpretation Psychologized. Mind & Language, 4(3), 161-185.

Goldman, A. I. (2006a). Simulating Minds. Oxford University Press.

Goldman, A. I. (2006b). Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading. Oxford University Press.

Goldman, A. I. (2009). Simulation Theory and Cognitive Neuroscience. In D. Murphy & M. Bishop (Eds.), Stich and His Critics.

Goldman, A. I., & Sripada, C. S. (2005). Simulationist models of face-based emotion recognition. Cognition, 94(3), 193-213.

Gordon, R. M. (1986). Folk Psychology as Simulation. Mind & Language, 1(2), 158-171.

Gordon, R. M. (1992). Reply to Stich and Nichols. Mind & Language, 7(12), 87-97.

Gordon, R. M. (1996). Radical Simulationism. In P. Carruthers & P. K. Smith (Eds.), Theories of Theories of Mind. Cambridge University Press.

Harris, P. L. (1992). From Simulation to Folk Psychology: The Case for Development. Mind & Language, 7(1-2), 120-144.

Haugeland, J. (1989). Artificial Intelligence: The Very Idea. Cambridge, Mass.: A Bradford Book.

Heal, J. (1996). Simulation and Cognitive Penetrability. Mind & Language, 11(1), 44-67.

Jackson, P. L., Meltzoff, A. N., & Decety, J. (2005). How do we perceive the pain of others? A window into the neural processes involved in empathy. NeuroImage, 24(3), 771-779.

Jeannerod, M. (2001). Neural Simulation of Action: A Unifying Mechanism for Motor Cognition. NeuroImage, 14(1), S103-S109.

Kahneman, D., & Tversky, A. (1981). The Simulation Heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), Judgment Under Uncertainty: Heuristics and Biases. Cambridge University Press.

Kosslyn, S. M., Pascual-Leone, A., Felician, O., Camposano, S., Keenan, J. P., L, W., ... Alpert. (1999). The Role of Area 17 in Visual Imagery: Convergent Evidence from PET and rTMS. Science, 284(5411), 167-170.

Lawrence, A. D., & Calder, A. J. (2004). Homologizing human emotions. In D. Evans & P. Cruse (Eds.), Emotion, Evolution, and Rationality (pp. 15-48). Oxford University Press.

Lawrence, A. D., Calder, A. J., McGowan, S. W., & Grasby, P. M. (2002). Selective disruption of the recognition of facial expressions of anger. Neuroreport, 13(6), 881-884.

Meltzoff, A. N., & Moore, M. K. (1977). Imitation of Facial and Manual Gestures by Human Neonates. Science, 198(4312), 75-78.

Meudell, P. R. (1971). Retrieval and representations in long-term memory. Psychonomic Science, 23(4), 295-296.

Nichols, S., & Stich, S. P. (2003). Mindreading. An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds. Oxford University Press.

Pellegrino, G. di, Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G.

(1992). Understanding motor events: a neurophysiological study. Experimental Brain Research, 91(1), 176-180.

Phillips, M. L., Young, A. W., Senior, C., Brammer, M., Andrew, C., Calder, A. J., David, A. S. (1997). A specific neural substrate for perceiving facial expressions of disgust. Nature, 389(6650), 495-498.

Pylyshyn, Z. W. (1973). What the minds eye tells the minds brain: A critique of mental imagery. Psychological Bulletin, 80(1), 1-24.

Ramsey, W. (2010). How not to build a hybrid: Simulation vs. fact-finding. Philosophical Psychology, 23(6), 775-795.

Rizzolatti, G., & Craighero, L. (2004). The Mirror-Neuron System. Annual Review of Neuroscience, 27(1), 169-192.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. Cognitive Brain Research, 3(2), 131-141.

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. Nature Reviews Neuroscience, 2(9), 661-670.

Rizzolatti, G., & Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. Nature Reviews Neuroscience, 11(4), 264-274.

Schwartz, D. L., & Black, T. (1999). Inferences through imagined actions: Knowing by simulated doing. Journal of Experimental Psychology: Learning, Memory, and Cognition, 25(1), 116-136.

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for Pain Involves the Affective but not Sensory Components of Pain. Science, 303(5661), 1157-1162.

Spaulding, S. (2012). Mirror neurons are not evidence for the Simulation Theory. Synthese, 189(3), 515-534.

Stich, S. P. (2009). Stich and His Critics. (M. Bishop & D. Murphy, Eds.). Blackwell.

Stich, S. P., & Nichols, S. (1995). Second thoughts on simulation. In Davies, Martin & T. Stone (Eds.), Mental Simulation: Evaluations and Applications.

Stich, S. P., & Nichols, S. (1997). Cognitive Penetrability, Rationality and Restricted Simulation. Mind & Language, 12(3–4), 297-326.

Wicker, B., Keysers, C., Plailly, J., Royet, J.-P., Gallese, V., & Rizzolatti, G. (2003). Both of Us Disgusted in My Insula: The Common Neural Basis of Seeing and Feeling Disgust. Neuron, 40(3), 655-664.

0.3 Simulation, Mirroring, and Neurological Similarity

Abstract

Simulation is an interesting hypothesis proposed to account for our ability of mental state attribution. However, in spite of its prominence, it remains fairly unclear how the theory, and the notion of simulation on which it depends, ought to be understood. This paper addresses the most promising sense of simulation, the neurological similarity, and argues that the processes involved in a class of celebrated simulation prototypes—that is, emotion recognition—do not demonstrate similarity in the sense simulation theorists contend.

0.3.1 Two Requirements for Similarity

One interesting, but unresolved, question in cognitive sciences concerns our ability of mental state attribution, often called mindreading. How is mindreading accomplished? According to an interesting view, the simulation theory (ST), we recognize and attribute mental states to others by using our own mind as a model for the other person's mental life. But, if P and P′ are two processes, what is required for P to qualify as a simulation of P′? Simulation theorists have been less than clear on this question, have used the term simulation in a vague and heterogeneous ways, for instance, as "imaginative identification" (Gordon, 1992), "imaginative transformation" (Gordon, 1996), placing oneself in the targets situation (Harris, 1992; Heal, 1996), and off-line simulation (Goldman, 1989). So, despite the theory's attractiveness to a group of advocates, it remains fairly unclear how the theory, and the notion of simulation on which it relies, ought to be understood.

In reaction to this situation, Stich and Nichols (Nichols & Stich, 2003; Stich, 2009; Stich & Nichols, 1997) have frequently remarked that the diversity among the processes to which simulation theorists have attached the label simulation "is so great that the term itself has become quite useless. It picks out no natural or theoretically interesting category" (Stich & Nichols, 1997, p. 299). However, Goldman, the most prominent defender of the simulation theory, resists this complaint and insists that "there is unity amid this diversity; simulation is still a natural and theoretically interesting category. Analogously, although there are many different atomic elements, the category "atomic element" is a natural and theoretically interesting category." (Goldman, 2006, p. 35). But Goldman "does not want to defend every application of the term 'simulation'that anybody has ever proposed" (Goldman, 2009, p. 138). So, he sets out his own definition of simulation:

> Process P simulates process P′ = df.
> (1) P duplicates, replicates, or resembles P′ in some significant respects (significant relative to the purposes or function of the task), and
> (2) in its (significant) duplication of P′, P fulfills one of its purposes or functions. (Goldman, 2006, p. 37)

This definition, although developed with care and in minute detail, is rather vague and offers no real help. It is fairly unclear what is precisely intended by duplication, or resemblance in (1), and what the terms "purpose" and "function" exactly designate in (2). Likely dimensions of similarity have already been discussed in the literature, including process similarity, concrete similarity, phenomenological, and functional similarity, all of them, however, confront serious problems (Fisher, 2006; Goldman, 1989, 2006; Spaulding, 2012). More recently, however, Goldman has argued that better prospects are found at the neurological level:

> At the neurological level, there may be better prospects for finding illuminating resemblances among processes... As I shall show, however, there is lots of evidence of neural resemblances and they greatly strengthen the case for simulation as a robust characteristic of both mindreading and other forms of social cognition. (Goldman, 2009b)

While the idea of neuronal similarity was first proposed in the motor domain (Gallese & Goldman, 1998), the most extensive pieces of evidence for neuronal similarity and mirror-based simulation, according to Goldman (2009b, 2009a), are found in studies on emotion recognition. Characterizing simulation in terms of neuronal similarity is supported by two sorts of evidence. First, brain imaging studies on normal participants show similar brain activation during both experience and recognition of emotions. Second, lesion studies demonstrate that damage to specific regions is accompanied by selective impairments in recognition of certain emotions.

Several findings from brain imaging and lesion studies are collected to show that the same neurological substrate underpins both experience and recognition of certain emotions. For instance, it is argued that the amygdala underpins both experience and recognition of fear, and the insula underpins experience and recognition of disgust. More importantly, it is argued that lesion studies demonstrate paired deficits (in experience and recognition) that are *selective*, that is, whereas damage to the neural region responsible for emotion E will impair mindreading E, it leaves intact recognition of other emotions E′, E″, and so on. Evidence from lesion studies shows that,

> these deficits were selective in the sense that patients impaired specifically in emotion X had no difficulty in recognizing emotion Y or Z but only in recognizing X.... One could also formulate the matter in terms of double, indeed triple, dissociations. Recognition of emotion X can be intact while recognition of Y is impaired, and recognition of Y can be intact while recognition of X is impaired; and so forth. (Goldman, 2009b, p. 145-146)

According to Goldman, the *selective* impairments together with *specific* activation of the regions in normal participants, demonstrate that experiencing an emotion and recognizing this same emotion depend on the same brain structure.

I contend that the similarity-based characterization of simulation hypothesis in terms of neurological similarity lays down two key requirements:

(1) that a brain region which is allegedly responsible for an emotion is *specifically* and *consistently* activated during both experience and recognition of the respective emotion, and

(2) that damage to a region responsible for experiencing a certain emotion would also impair recognition of that emotion.

Condition (1) requires specific and consistent activation. Specificity requires that specific brain regions (e.g. the amygdala) be preferentially active for tokens of one, and only one, emotion category (e.g. fear). Consistency requires that a region active for an emotion category to be active for every token of that category. Condition (2) requires paired deficits in both experience and recognition of an emotion. Let us call (1) and (2) the (neurological) similarity requirements. In what follows, I will discuss the two basic emotions of fear and disgust, where simulation theory has its best evidence for neurological similarity, and argue that the processes involved in emotion recognition fail to satisfy the similarity requirements.

Before moving on I want to preempt an objection one could make on behalf of Goldman's account. In addition to his definition of simulation, Goldman draws an important distinction between *successful* and *attempted* simulation. Attempted mental simulation is defined as follow:

> Process P is an attempted mental simulation of process P′ if P and P′ are both mental processes, and P is executed with the *aim* of duplicating or matching P′ in some significant respects. (Goldman, 2006, p. 38)

Simulation is either successful or attempted, and not all attempted simula-

tions are accurate or successful. Neurological similarity between the simulating system and target is expected only in cases of successful simulation. Thus, simulation is not committed to the idea that mindreading will always, or even usually, show neurological similarity.

This objection is unfounded. Goldman has distinguished between low-level and high-level mindreading. High-level mindreading is the partly voluntary and to some degree conscious process of detecting propositional states guided by imagination, and low-level mindreading is the automatic and unconscious detection of emotional states (Goldman, 2006, p. 43). Now, whereas high-level mindreading requires only attempted, not successful, simulation, low-level mindreading is subserved by more primitive, automatic and mirror mechanisms which requires genuine neurological resemblances. Thus, although mistaken mental state attributions at high-level mindreading might be accommodated as attempted simulation, low-level mindreading is predicated "on genuine resemblances between states of the attributor and the target. The case for high-level mindreading, by contrast, rests on the ostensible purpose or function of E-imagination, not on the regular achievement of faithful reproductions" (Goldman, 2006, p. 150). A process counts as "attempted mental simulation" only if the process is executed with the *aim* of duplicating the processes of the target (see the definition above). However, the term aim does not appear in Goldman's account of low-level mindreading. And, it hardly makes sense that a neuronal process is executed with the aim of duplicating a neuronal target process.

With this observation in mind, let us see how low-level simulational account of fear and disgust meets the similarity requirements of (1) and (2).

0.3.2 Simulation and Neurological Similarity

Beginning with fear, we must ask, is the amygdala specifically a fear processor? As Sander and colleagues remark (Sander, Grafman, & Zalla, 2003), to show the specificity it needs to be demonstrated that the difference obtained in the amygdala activation for fear-inducing stimuli (versus neutral stimuli) cannot be obtained when comparing amygdala activation for other non-fear-related stimuli. However, most brain imaging studies show amygdala activation for several emotions. For instance, Blair et al. (Blair, Morris, Frith, Perrett, & Dolan, 1999) scanned thirteen normal subjects when subjects viewed images of faces expressing varying degrees of sadness and anger while performing a sex discrimination task (whether the person whose face they saw was male or female). The results show enhanced activity of the left amygdala associated with increasing intensity of sad and angry facial expressions. Similarly, Whalen et al. (2001) found amygdala activation for

angry expressions, and results from an fMRI study by Siegle et al. (Siegle, Steinhauer, Thase, Stenger, & Carter, 2002) show amygdala activation to negative, but not specifically fear-related, information in depressed individuals.

In addition, other findings form brain imaging studies have found a broader role for the amygdala than fear processing. Breiter et al. (1996) found amygdala activation for happy versus neutral faces. In a brain imaging study (Garavan, Pendergrass, Ross, Stein, & Risinger, 2001), subjects viewed pictures that varied in emotional content (positive versus negative valence) while undergoing fMRI scanning. The results showed significant amygdala activation for both positively and negatively valenced stimuli. Also, Wright and colleagues (Wright, Martis, Shin, Fischer, & Rauch, 2002) evaluated human brain responses to simple drawings of emotional and neutral facial expressions. Significantly-increased fMRI signal was found in the amygdala in response to angry and happy schematic faces. Some studies (Kim, Somerville, Johnstone, Alexander, & Whalen, 2003; Wang, McCarthy, Song, & Labar, 2005) suggest amygdala's role in processing facial expressions of surprise and sadness. Other findings show amygdala response across all emotional expressions. For instance, Yang et al. (2002) examined the amygdala in response to the perception of happy, angry, sad and fearful facial expressions compared to neutral expressions. The results demonstrate that all four facial expressions, including happy faces, were associated with reliable bilateral activation of the amygdala. Similarly, in a study by Winston et al. (Winston, ODoherty, & Dolan, 2003), subjects viewed morphed emotional faces displaying low and high intensities of disgust, fear, happiness or sadness under two different task conditions. The amygdala responded to high-intensity expressions of all the four basic emotions, suggesting that the amygdala is involved in perceptual processing of a range of emotions, rather than fear only. In a brain imaging study by Fitzgerald and colleagues (Fitzgerald, Angstadt, Jelsone, Nathan, & Phan, 2006), 20 subjects viewed photographs displaying fearful, disgusted, angry, sad, neutral and happy facial expressions. The left amygdala was activated by each condition separately (emotional or non-emotional), and its response was not selective for any particular emotion category.

In addition, the link between amygdala activation and fear-inducing stimuli has not been entirely consistent. Schienle et al. (A. Schienle et al., 2002) and Stark et al. (Stark et al., 2003) obtained no amygdala activation. This may suggest that the amygdala is not even necessary in fear processing. Further, recent findings by Tsuchiya et al. (Tsuchiya, Moradi, Felsen, Yamazaki, & Adolphs, 2009) and Piech et al. (Piech et al., 2010) on a patient SM with bilateral amygdala damage show that SM's speed performance on a rapid detection task of fear-related stimuli was completely normal. This result, as

the authors remark, suggests that the amygdala is not essential at least for the early stages of fear processing.

I argued that amygdala is neither specifically nor consistently activated during recognition of fear. That is, fear recognition does not satisfy (1). Fear recognition does not satisfy (2) either; fear recognition is preserved despite unilateral or bilateral amygdala damage. Hamann et al. (1996) examine two patients, EP and GP, with complete bilateral amygdala damage and additional lesions in temporal lobe structures. EP and GP were tested twice, once using the exact material and procedures reported by Adolphs et al., (1994), where they had previously found poor performance in a fear recognition task in a patient with bilateral amygdala damage, and again tested the patients in a slightly different version of the same experiment. The patients, despite bilateral amygdala damage, appeared to be unimpaired in recognition of fear or any other emotion category.

Besides, that one and the same neural substrate is not implicated in both experience and recognition of fear is further confirmed by observing that damage to regions other than the amygdala can impair fear recognition. For instance, on a face emotion recognition task, Adolphs and colleagues (Adolphs, Damasio, Tranel, & Damasio, 1996) examined 37 subjects with focal brain damage and compared their performances to the mean performance of 15 normal controls. First, the study found no recognition impairment in subjects with lesions restricted to left hemisphere. However, damage to the right hemisphere was associated with impairments in emotion recognition. Second, damage to the right anterior infracalcarine cortex was associated with impairments in fear recognition. Taken together, the above results show that fear recognition fails to satisfy both (1) and (2).

The next emotion is disgust. Does disgust recognition satisfy the similarity conditions of (1) and (2)? It does not satisfy (1). To show that the insula is specifically a disgust processor, it must be demonstrated that the difference in activation obtained for disgust cannot be obtained for other emotion categories. However, the brain imaging studies by Philips et al. (1997) and Wicker et al. (2003), on which the simulation account of disgust recognition relies, examine the insula activation exclusively with respect to disgust; they compare the insula activation for disgust relative to neutral faces only. To find the desired specificity, insula activation must be compared with respect to other emotions as well. Several studies, however, show that upon such comparisons, one can hardly conclude that insula is specifically a disgust processor. For instance, Schienle et al. (2002) presented subjects with pictures displaying a wide variety of different disgust and fear elicitors. The results show insula activation during the fear condition, and amygdala activation during the disgust condition. As the authors remark, the finding

accords with the notion of the insula as a region involved in affective tasks without focusing on any specific emotion, a view which is also similar to Damasios (Damasio et al., 2000) conception of the insula as part of a central circuit concerned with monitoring emotional states in general.

In addition, although the study by Phillips et al. (1998) showed insula activation in response to disgust, they also found activation in the insula for fear, and in the amygdala for disgust. In another study, Stark et al. (2003) also found amygdala activation during the disgust condition. Besides, similar brain structures were activated when Stark and colleagues contrasted the disgusting and fear-inducing pictures with the affectively neutral pictures, a finding which suggests the idea that fear and disgust are processed in similar brain structures.

In addition, the link between insula activation and disgust has not been entirely consistent. For instance, Phillips et al. (1998) didn't find the insula or basal ganglia activation in response to auditory disgust stimuli. Furthermore, Schienle et al. (Schienle, Schfer, Stark, Walter, & Vaitl, 2005) analyzed data from 63 subjects across four studies to see if the insula, the amygdala, the orbitofrontal cortex and the medial prefrontal cortex would be involved in disgust processing. Whereas subjects experienced intense feelings of disgust (based on a self-report questionnaire), the study found no insular activation during the disgust conditions.

Moreover, studies show regions other than the amygdala that are involved in disgust processing. Evidence from several studies (Gorno-Tempini et al., 2001; Schienle et al., 2006, 2005; Winston et al., 2003) show amygdala activation in response to disgust. For instance, in a study by Schienle et al. (2006), two types of disgust elicitors (pictures of contamination and humiliation) were compared with fear-relevant and neutral (scene) stimuli. The results show that both of the disgust conditions elicited activations in the amygdala, the occipitotemporal cortex, and the orbitofrontal cortex, but no significant activation in the insula.

I argued that the insula neither specifically nor consistently responds during disgust recognition. That is, disgust recognition doesn't satisfy (1). I now argue that it doesn't satisfy (2) either. It is notable that patients with focal brain damage to the insula or basal ganglia are rare. For a decade, NK was the only patient with insula and basal ganglia damage whose score on a disgust experience questionnaire and performance on a disgust recognition task was lower than controls (Calder, Keane, Manes, Antoun, & Young, 2000). However, it is notable that NK's poor performance involved processing of other emotions as well. For instance, during tests of non-verbal emotional sounds, NK showed a deficit in recognizing surprise and incorrectly labeled disgust as fear and anger. When interpreting JACFEE facial expressions,

NK's recognition of contempt was impaired, and he miscategorized disgust as anger and contempt. In addition, NK incorrectly categorized disgust facial expression as anger when the Ekman and Friesen faces were used.

In addition, more recently Straube et al. (2010) reported a patient MK with a lesion comparable to NK in the insula and basal ganglia. If the insula and basal ganglia are reliably involved in disgust processing, MK should show at least some deficits in the processing of disgust stimuli. To examine MK, Straube and colleagues used tests and methods similar to those used by Calder et al. (2000) in their study of NK. Contrary to Calder et al.'s findings, none of the tests by Straube and colleagues on MK revealed a deficit in disgust processing. Compared to healthy controls, MK showed no impairments in the recognition or experience of disgust, nor any impairment in the recognition or experience of other emotions. This finding is corroborated by a more recent study by Couto et al. (2013) on a patient GG with focal insula damage. Similar to MK, GG showed no impairment in emotion recognition. The two studies show preserved disgust experience and recognition despite exclusive focal damage to the basal ganglia and insular cortex.

Because focal brain damage to the insula and basal ganglia are very infrequent, other types of lesion studies might be more illuminating here. Patients with Huntington's disease (HD), a neurodegenerative disorder that affects the basal ganglia and insula, are of primary interest in this area. Indeed, the primary evidence for the idea that disgust might be processed in a specific brain structure came from findings in patients with Huntingtons disease, when Sprengelmeyer and colleagues (R. Sprengelmeyer et al., 1996) found HD patients showing severe impairments during the facial and vocal motion recognition tests. However, even in this early study, the average rate at which disgust was detected was below the next most badly affected emotion, fear. In addition, the patients had severe problems in discriminating fear from anger. In general, results from this study show impairments in recognition of most emotions, with some emotions impaired more than others. In a follow-up study, Sprengelmeyer et al. (1997) examined emotion recognition at an individual level in two HD patients, HL and UJ. The results show severe impairments in the recognition of disgust and fear, but not disgust only, and one of the patients, UJ, had problems involving the misrecognition of fear as anger.

Several other findings show that disgust recognition is preserved despite damage to the basal ganglia or the insula. Milders et al. (Milders, Crawford, Lamb, & Simpson, 2003) compared the performance of HD patients and gene-carriers of HD on two sets of emotion recognition tests and found that HD patients were impaired at recognizing several expressions, including sadness, anger, disgust and fear, compared to healthy controls and asymp-

tomatic gene-carriers. Interestingly, there was no indication of a selective impairment in disgust recognition. Indeed, further testing on selective impairment revealed that the patients were in fact significantly more impaired on other negative emotions, for instance fear, than on disgust. The result was corroborated by three more emotion recognition studies—one of them on 475 HD patients—that found decline in recognition of all negative emotions, including sadness, fear, disgust, anger, and surprise (Ille et al., 2011; Johnson et al., 2007; Snowden et al., 2008). Taken together, the above results show that disgust recognition fails to satisfy both (1) and (2).

I have shown that the processes involved in experience and recognition of fear and anger, where simulation theorists have found the best evidence for neuronal similarity, do now show neural resemblance in the sense required by the similarity conditions. The result holds not only for fear and anger but also for other emotions as well, because emotions and sensations are processed not in specific brain structures as it is claimed, but in an integrative system involving multiple cross-regional interactions in the brain. This result faces simulation theorists with a conceptual question simulationists have been struggling with for more than two decades, this time, however, at low-level mindreading. If emotions are processed in an integrative and distributed system, how can simulation (at low-level) be characterized in terms of neuronal similarity? Unless the question is answered, we do not know how and precisely in what sense emotion recognition is simulational.

0.3.3 Bibliography

Adolphs, R., Damasio, H., Tranel, D., & Damasio, A. R. (1996). Cortical systems for the recognition of emotion in facial expressions. The Journal of Neuroscience, 16, 7678-7687.

Blair, R. J. R., Morris, J. S., Frith, C. D., Perrett, D. I., & Dolan, R. J. (1999). Dissociable neural responses to facial expressions of sadness and anger. Brain, 122, 883-893.

Breiter, H. C., Etcoff, N. L., Whalen, P. J., Kennedy, W. A., Rauch, S. L., Buckner, R. L., ... Rosen, B. R. (1996). Response and habituation of the human amygdala during visual processing of facial expression. Neuron, 17, 875-887.

Calder, A. J., Keane, J., Manes, F., Antoun, N., & Young, A. W. (2000). Impaired recognition and experience of disgust following brain injury. Nature Neuroscience, 3, 1077-1078.

Couto, B., Sedeo, L., Sposato, L. A., Sigman, M., Riccio, P. M., Salles, A., ... Ibanez, A. (2013). Insular networks for emotional processing and social cognition: Comparison of two case reports with either cortical or subcortical involvement. Cortex; a Journal Devoted to the Study of the Nervous System and Behavior, 49, 1420-1434.

Damasio, A. R., Grabowski, T. J., Bechara, A., Damasio, H., Ponto, L. L. B., Parvizi, J., & Hichwa, R. D. (2000). Subcortical and cortical brain activity during the feeling of self-generated emotions. Nature Neuroscience, 3, 1049-1056.

Fisher, J. C. (2006). Does Simulation Theory Really Involve Simulation? Philosophical Psychology, 19(4), 417-432.

Fitzgerald, D. A., Angstadt, M., Jelsone, L. M., Nathan, P. J., & Phan, K. L. (2006). Beyond threat: Amygdala reactivity across multiple expressions of facial affect. NeuroImage, 30, 1441-1448.

Gallese, V., & Goldman, A. I. (1998). Mirror neurons and the simulation theory of mindreading. Trends in Cognitive Sciences, 2, 493-501.

Garavan, H., Pendergrass, J. C., Ross, T. J., Stein, E. A., & Risinger, R. C. (2001). Amygdala response to both positively and negatively valenced stimuli. Neuroreport, 12, 2779-2783.

Goldman, A. I. (1989). Interpretation Psychologized. Mind & Language, 4(3), 161-185.

Goldman, A. I. (2006). Simulating Minds: The Philosophy, Psychology, and

Neuroscience of Mindreading. Oxford University Press.

Goldman, A. I. (2009a). Mirroring, Simulating and Mindreading. Mind & Language, 24(2), 235-252.

Goldman, A. I. (2009b). Simulation Theory and Cognitive Neuroscience. In D. Murphy & M. Bishop (Eds.), Stich and His Critics.

Gordon, R. M. (1992). Reply to Stich and Nichols. Mind & Language, 7(12), 87-97.

Gordon, R. M. (1996). "Radical" Simulationism. In P. Carruthers & P. K. Smith (Eds.), Theories of Theories of Mind. Cambridge University Press.

Gorno-Tempini, M. L., Pradelli, S., Serafini, M., Pagnoni, G., Baraldi, P., Porro, C., ... Nichelli, P. (2001). Explicit and Incidental Facial Expression Processing: An fMRI Study. NeuroImage, 14(2), 465-473.

Hamann, S. B., Stefanacci, L., Squire, L. R., Adolphs, R., Tranel, D., Damasio, H., & Damasio, A. (1996). Recognizing facial emotion. Nature, 379(6565), 497.

Harris, P. L. (1992). From Simulation to Folk Psychology: The Case for Development. Mind & Language, 7(12), 120-144.

Heal, J. (1996). Simulation and Cognitive Penetrability. Mind & Language, 11(1), 44-67.

Ille, R., Holl, A. K., Kapfhammer, H.-P., Reisinger, K., Schfer, A., & Schienle, A. (2011). Emotion recognition and experience in Huntingtons disease: is there a differential impairment? Psychiatry Research, 188(3), 377-382.

Johnson, S. A., Stout, J. C., Solomon, A. C., Langbehn, D. R., Aylward, E. H., Cruce, C. B., ... Group, the P.-H. I. of the H. S. (2007). Beyond disgust: impaired recognition of negative emotions prior to diagnosis in Huntingtons disease. Brain, 130(7), 1732-1744.

Kim, H., Somerville, L. H., Johnstone, T., Alexander, A. L., & Whalen, P. J. (2003). Inverse amygdala and medial prefrontal cortex responses to surprised faces. Neuroreport, 14(18), 2317-2322.

Milders, M., Crawford, J. R., Lamb, A., & Simpson, S. A. (2003). Differential deficits in expression recognition in gene-carriers and patients with Huntington's disease. Neuropsychologia, 41(11), 1484-1492.

Nichols, S., & Stich, S. P. (2003). Mindreading. An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds. Oxford University Press.

Phillips, M. L., Young, A. W., Scott, S. K., Calder, A. J., Andrew, C., Giampietro, V., ... Gray, J. A. (1998). Neural responses to facial and vocal expressions of fear and disgust. Proceedings of the Royal Society of London B: Biological Sciences, 265(1408), 1809-1817.

Piech, R. M., McHugo, M., Smith, S. D., Dukic, M. S., Meer, J. V. D., Abou-Khalil, B., & Zald, D. H. (2010). Fear-enhanced visual search persists after amygdala lesions. Neuropsychologia, 48(12), 3430-3435.

Sander, D., Grafman, J., & Zalla, T. (2003). The human amygdala: an evolved system for relevance detection. Reviews in the Neurosciences, 14(4), 303-316.

Schienle, A., Schfer, A., Hermann, A., Walter, B., Stark, R., & Vaitl, D. (2006). fMRI responses to pictures of mutilation and contamination. Neuroscience Letters, 393(2–3), 174-178.

Schienle, A., Schfer, A., Stark, R., Walter, B., & Vaitl, D. (2005). Relationship between disgust sensitivity, trait anxiety and brain activity during disgust induction. Neuropsychobiology, 51(2), 86-92.

Schienle, A., Stark, R., Walter, B., Blecker, C., Ott, U., Kirsch, P., ... Vaitl, D. (2002). The insula is not specifically involved in disgust processing: an fMRI study. Neuroreport, 13(16), 2023-2026.

Siegle, G. J., Steinhauer, S. R., Thase, M. E., Stenger, V. A., & Carter, C. S. (2002). Can't shake that feeling: event-related fMRI assessment of sustained amygdala activity in response to emotional information in depressed individuals. Biological Psychiatry, 51(9), 693-707.

Snowden, J. S., Austin, N. A., Sembi, S., Thompson, J. C., Craufurd, D., & Neary, D. (2008). Emotion recognition in Huntingtons disease and frontotemporal dementia. Neuropsychologia, 46(11), 2638-2649.

Spaulding, S. (2012). Mirror neurons are not evidence for the Simulation Theory. Synthese, 189(3), 515–534.

Sprengelmeyer, R., Young, A. W., Calder, A. J., Karnat, A., Lange, H., Hmberg, V., ... Rowland, D. (1996). Loss of disgust. Perception of faces and emotions in Huntingtons disease. Brain: A Journal of Neurology, 119 (Pt 5), 1647-1665.

Sprengelmeyer, R., Young, A. W., Sprengelmeyer, A., Calder, A. J., Rowland, D., Perrett, D., ... Lange, H. (1997). Recognition of Facial Expressions: Selective Impairment of Specific Emotions in Huntingtons Disease. Cognitive Neuropsychology, 14(6), 839-879.

Stark, R., Schienle, A., Walter, B., Kirsch, P., Sammer, G., Ott, U., ... Vaitl, D. (2003). Hemodynamic responses to fear and disgust-inducing pictures: an fMRI study. International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology, 50(3), 225-234.

Stich, S. P. (2009). Stich and His Critics. (M. Bishop & D. Murphy, Eds.). Blackwell.

Stich, S. P., & Nichols, S. (1997). Cognitive Penetrability, Rationality and Restricted Simulation. Mind & Language, 12(3-4), 297–326.

Straube, T., Weisbrod, A., Schmidt, S., Raschdorf, C., Preul, C., Mentzel, H.-J., & Miltner, W. H. R. (2010). No impairment of recognition and experience of disgust in a patient with a right-hemispheric lesion of the insula and basal ganglia. Neuropsychologia, 48(6), 1735-1741.

Tsuchiya, N., Moradi, F., Felsen, C., Yamazaki, M., & Adolphs, R. (2009). Intact rapid detection of fearful faces in the absence of the amygdala. Nature Neuroscience, 12(10), 1224-1225.

Wang, L., McCarthy, G., Song, A. W., & Labar, K. S. (2005). Amygdala activation to sad pictures during high-field (4 tesla) functional magnetic resonance imaging. Emotion (Washington, D.C.), 5(1), 12-22.

Whalen, P. J., Shin, L. M., McInerney, S. C., Fischer, H., Wright, C. I., & Rauch, S. L. (2001). A functional MRI study of human amygdala responses to facial expressions of fear versus anger. Emotion, 1(1), 70-83.

Winston, J. S., ODoherty, J., & Dolan, R. J. (2003). Common and distinct neural responses during direct and incidental processing of multiple facial emotions. NeuroImage, 20(1), 84-97.

Wright, C. I., Martis, B., Shin, L. M., Fischer, H., & Rauch, S. L. (2002). Enhanced amygdala responses to emotional versus neutral schematic facial expressions. Neuroreport, 13(6), 785-790.

Yang, T. T., Menon, V., Eliez, S., Blasey, C., White, C. D., Reid, A. J., ... Reiss, A. L. (2002). Amygdalar activation associated with positive and negative facial expressions. Neuroreport, 13(14), 1737-1741.

0.4 Why Emotion Recognition Is Not Simulational

Abstract

According to a dominant interpretation of the simulation hypothesis, in recognizing an emotion we use the same neural processes used in experiencing that emotion. This paper argues that the view is fundamentally misguided. I will examine the simulational arguments for the three basic emotions of fear, disgust, and anger and argue that the simulational account relies strongly on a narrow sense of emotion processing which hardly squares with evidence on how, in fact, emotion recognition is processed. I contend that the current body of empirical evidence suggests that emotion recognition is processed in an integrative system involving multiple cross-regional interactions in the brain, a view which squares with understanding emotion recognition as an information-rich, rather than simulational, process. In the final section, I discuss possible objections.

0.4.1 Theory-Theory vs. Simulation Theory

An individual shows the mindreading ability if he or she attributes mental states to self or others. What kinds of processes and mechanisms underlie mental state attribution in humans? The question has been central over the last two decades and has given rise to the two major positions of theory theory (TT) and simulation theory (ST). According to TT, mindreading is guided and executed by a theory of mind, where the term "theory" is variously construed as law-like generalizations (Churchland, 1981), internally represented knowledge structure (Stich & Nichols, 1992), or model theories (Godfrey-Smith, 2005; Maibom, 2007, 2009). In more general terms, TT is an account on which mindreading is understood as an information-rich process. A typical information-rich style explanation posits a set of mental representations that contain information about external stimuli, mental states, observable behavior, and the relation among them. However, the alternative hypothesis, ST, holds we understand others using our own mind/brain as a model for the other persons mental life. The mental representations postulated by TT theorists are explanatorily redundant because, according to ST, mindreading is achieved by the operation of purely observation/execution matching mechanisms (Goldman, 2006; Gordon, 1986; Heal, 1986). Besides, several theorists have developed accounts which take elements of both theory and simulation (Botterill & Carruthers, 1999; Heal, 1995; Nichols & Stich, 2003; Perner & Kuhberger, 2005).

More recently, simulation theorists have argued that findings from cognitive neuroscience, specifically the discovery of mirror neurons

(Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992; Rizzolatti, Fadiga, Gallese, & Fogassi, 1996), are evidence in support of their position. Mirror neurons are a specific class of neurons that discharge both when a monkey performs an action and when it observes the same action by another individual. In addition to motor actions, mirror neurons and mirror processes have been discovered across different domains, including the sensation, perception, and emotional response to pain (Jackson, Meltzoff, & Decety, 2005; Singer et al., 2004), touch (Keysers et al., 2004), and disgust (Wicker et al., 2003), among others. In each case, the brain in its observation mode behaves as if it mirrors the activity of the brain in its execution/experience mode.[5] While the term "mirroring" technically only refers to an overlap in neuronal discharge between different modes of endogenous and exogenous activation, it has been speculated that the processes underlying mirroring play a significant role in some cognitive events, such as imitation (Iacoboni et al., 1999; Rizzolatti, Fogassi, & Gallese, 2001), empathy (Decety & Jackson, 2004; Gallese, 2001), and mindreading (Gallese & Goldman, 1998; Goldman, 2006; Goldman & Sripada, 2005).

Goldman, the most prominent defender of the simulation theory, has distinguished between low- and high-level mindreading. High-level mindreading is the partly voluntary and to some degree conscious process of detecting propositional states guided by an imagination-driven form of simulation, whereas low-level mindreading is the automatic and unconscious detection of emotional states by mirror-based simulation. (Goldman, 2006, p. 43). For low-level mindreading, my focus in this paper, the best and most extensive pieces of evidence are found in studies on emotion recognition (Goldman, 2009a, pp. 243–244). So, if we take simulation as a process P that duplicates or resembles another process P′ in the target [6] (Goldman, 2006, p. 37), and understand the relevant respect of resemblance between P and P′ in terms of neurological similarity [7], what would be the possible simulational account of a face-based emotion recognition with respect to some emotion E? On the ST account, people use the same neural system in making a recognition judgment as is used in experiencing emotion E (Goldman, 2009b, p. 147). That being so, ST predicts that damage to a brain region responsible for the production of an emotion will also impair

[5]For a critique of evidence for mirror neurons, see, for example, Dinstein, Thomas, Behrmann, and Heeger (2008) and Hickok (2009, 2014).

[6]Note that simulation can be intra-personal or inter-personal. Here we are discussing third person mindreading and inter-personal simulation.

[7]P and P′ might resemble each other in at least three different respects, phenomenological, functional, and neurological, but the neurological respect is the most promising one for low-level mindreading. For a discussion, see Goldman (2009b).

recognition of that emotion. But is there any evidence to support that emotion experience and recognition co-occur?

Goldman recounts a number of lesion and brain imaging studies in which the pattern (in its strongest form) emerges in at least three basic emotions of fear, disgust, and anger. Goldman shows that studies on both brain-damaged patients and healthy subjects tell us that the bilateral amygdala is involved both in the experience and recognition of fear, the insula and basal ganglia are involved in the experience and recognition of disgust, and dopamine levels are involved in the experience and recognition of anger. The evidence, Goldman argues, reflects a systematic relationship between emotion experience and recognition such that for an emotion E_1, there is a region R_1 that instantiates both experience and recognition of E_1, and someone impaired in experiencing E_1 would be also impaired in recognizing E_1. Thus, simulation is the best explanation for the fact that experiencing an emotion and recognizing this same emotion depend on the same brain structure (Goldman, 2006, 2008, 2009b, 2012; Goldman & Jordan, 2013; Goldman & Sripada, 2005).

But what would be the possible TT account of emotion recognition? The paired deficits can also be explained under the TT hypothesis, so long as one of two assumptions holds, both of which, however, are put into question by Goldman. If, on TT account, emotion recognition depends on mental representations that map particular expressions to certain emotions, then one option for TT to explain the impairments is to assume that experience of an emotion and the representations involved in recognition of that emotion occur in the same brain structure. A neural region might be used as a substrate. It could be the case that a neural region is used both as a neural substrate for instantiating an emotion and as a substrate for representing information about that emotion but, Goldman argues, "why should conceptual representations of fear occur in the same region [e.g., the amygdala] that underlies fear experience?" (Goldman & Sripada, 2005, p. 199). An alternative option for TT theorists is to think of a brain region as a module dedicated to emotion processing in general. A dedicated region, according to Goldman (Goldman, 2006; Goldman & Sripada, 2005), does not seem to be an option because damage to this region would impair recognition of all emotions. If so, how would TT explain impairment of a specific emotion? I will discuss Goldman's argument later but for now the important point is that Goldman's simulational account and his argument against TT is based on the evidence he presents for selective impairments. Goldman argues that whereas damage to the neural region responsible for emotion E will impair mindreading E, it leaves intact recognition of other

emotions E′, E″, and so on. Evidence from lesion studies shows that,

> these deficits were selective in the sense that patients impaired specifically in emotion X had no difficulty in recognizing emotion Y or Z but only in recognizing X . . . One could also formulate the matter in terms of double, indeed triple, dissociations. Recognition of emotion X can be intact while recognition of Y is impaired, and recognition of Y can be intact while recognition of X is impaired; and so forth. (Goldman, 2009b, pp. 145-146)

The selective impairments, according to Goldman, demonstrate that simulation theory is the best explanation of the fact that experiencing an emotion and recognizing this same emotion depend on the same brain structure. Thus, ST is an explanatorily richer hypothesis than TT, once evidence from cognitive neuroscience is taken into account.

In this paper, I will show that Goldman's simulational account of emotion recognition, his arguments against TT, and the idea of selective impairment behind them, depend on a narrow sense of emotion processing which involves:

(1) *Specificity*: that specific brain regions (e.g., the amygdala) be preferentially active for tokens of one, and only one, emotion category (e.g., fear).

(2) *Consistency*: that a region active for an emotion category is active for every token of that category.

(3) *Selectivity*: that impairments are selective in the sense that deficit in a brain region involved in an emotion will impair recognition of its respective, and only its respective, emotion category.

I will argue that this conception of emotion recognition is fundamentally flawed. I will discuss a body of evidence from brain imaging and lesion studies for three basic emotions of fear, disgust, and anger, where Goldman's simulational account has its strongest evidence, and argue that (1) specific brain regions are involved in instantiating different emotion categories, (2) that for a region that is allegedly specific to an emotion, inconsistent activations are abundant, and (3) patients with specific brain lesions reveal several, rather than selective, impairments. The emerging picture suggests that emotions are processed in multiple integrated and distributed brain structures. This outcome has two important implications. First, it shows that Goldman's account of emotion recognition hardly squares with evidence on how in fact emotion recognition is processed. Second, it defuses Goldman's argument against possible information-rich explanations of emotion recognition.

Simulation has been hypothesized for several emotions and sensations, but here I only discuss fear, disgust, and anger, which provide

the best evidence for the low-level simulation theory. I begin each section with a brief review of Goldman's argument for each emotion, followed by my discussion of that emotion.

0.4.2 Fear and the Amygdala

The first piece of evidence for a selectively paired deficit comes from studies on fear. In an early study, Adolphs, Tranel, Damasio, and Damasio (1994) studied a patient SM with bilateral amygdala damage. When SM was tested in a face-based emotion recognition task,[8] her rating of fearful faces correlated less with normal ratings than did those of any brain-damaged controls. In addition, other studies (Adolphs & Tranel, 2000) suggest that SM is also abnormal in her experience of fear. A different patient, NM, with bilateral amygdala damage was studied by Sprengelmeyer and colleagues (1999). Similar to SM, NM was also abnormal in his experience and recognition of fear. Goldman maintains that NM was impaired *only* in fear recognition and that other neuropsychological studies on the amygdala are broadly consistent with these findings of selectively paired deficits. The findings altogether, Goldman argues, strongly suggest that normal people use one and the same neuronal region, the amygdala, in recognizing fear as they would in experiencing fear. Therefore, fear recognition is simulational. (Goldman, 2006; Goldman & Sripada, 2005).

However, this analysis leaves out a large body of evidence that demonstrates that fear recognition does not, in fact, work in the way that Goldman's account predicts. First, although a deficit in recognition of fearful expressions is often observed in patients with amygdala damage, studies show that the deficit usually impairs recognition of other emotions as well. For instance, SM rated not only fear but also expressions of anger and surprise as less intense than did any of the brain-damaged controls (Adolphs et al., 1994). In a subsequent study on subjects with unilateral and bilateral amygdala damage, Adolphs and colleagues (1995) found SM rating surprised and angry faces as signaling less intense expressions of surprise and anger than did any of the controls, and all subjects with left amygdala damage rated disgust and sadness expressions as less intense than did subjects with right amygdala damage and brain-damaged controls. In other cases with bilateral amygdala damage, NM showed differing experience of anger and also difficulty in recognition of sadness from facial expressions

[8]Subjects were shown facial expressions of six basic emotions, as well as neutral faces, and asked to rate each face according to several emotional adjectives (Adolphs et al., 1994).

(Sprengelmeyer et al., 1999), and the patient DR showed impaired recognition of anger and (to a less extent) disgust (Scott et al., 1997). NM reported that he rarely experienced fear and anger when asked about the occurrence of fear and anger in everyday situations. Results from a subsequent collaborative study on nine subjects with bilateral amygdala damage showed that the subjects all gave abnormally low rating scores for most negative emotions (Adolphs et al., 1999).

The view that the amygdala is not specifically a fear processor can further be supported by evidence from brain imaging studies. As Sander, Grafman, and Zalla (2003) remark, to show that the amygdala is specifically involved in fear processing, it needs to be demonstrated that the difference obtained in the amygdala activation for fear-inducing stimuli (versus neutral stimuli) cannot be obtained when comparing amygdala activation for other non-fear-related stimuli. However, most brain imaging studies show amygdala activation for several emotions. In one study, (Blair, Morris, Frith, Perrett, & Dolan, 1999) the authors scanned 13 normal subjects while subjects viewed images of faces expressing varying degrees of sadness and anger while performing a sex discrimination task (responding as to whether the person whose face they saw was male or female). Blair and colleagues found enhanced activity in the left amygdala associated with increasing intensity of sad and angry facial expressions. In addition, Whalen and colleagues (2001) found amygdala activation for angry expressions, and results from an fMRI study (Siegle, Steinhauer, Thase, Stenger, & Carter, 2002) show amygdala activation to negative, but not specifically fear-related, information in depressed individuals.

Second, several studies show preserved fear recognition despite unilateral or bilateral amygdala damage. Hamann and colleagues (1996) report data from two patients, EP and GP, with complete bilateral lesions of the amygdala and additional temporal lobe structures. EP and GP were tested twice, once using the exact material and procedures reported by Adolphs and colleagues (1994), and again in a slightly different version of the same experiment. Surprisingly, EP and GP rated normally the same facial expressions that SM (see above) rated abnormally. The patients, despite bilateral amygdala damage, appeared to be unimpaired in recognition of fear or any other emotion category.

Third, studies show that damage to regions other than the amygdala can also result in impairment of fear recognition. For instance, Adolphs and colleagues (1996) examined 37 subjects with focal brain damage on a face recognition task and compared their performances to the mean performance of 15 normal controls. Two findings in this study are remarkable. First, the study found no recognition impairment in subjects with lesions restricted to left hemisphere. However,

damage to the right hemisphere was associated with impairments in emotion recognition. Second, damage to the right anterior infracalcarine cortex was associated with impairments in fear recognition. Brain imaging studies also show activation of regions other than the amygdala during the fear condition. Schienle and colleagues (2002) found insula activation during the fear conditions. Besides, examination of emotion processing in patients with obsessive compulsive disorder shows that fear processing involves multiple regions, including the orbitofrontal cortex and the insula (Schienle, Schfer, Stark, Walter, & Vaitl, 2005a).

Moreover, the link between amygdala activation and fear-inducing stimuli has not been entirely consistent. Schienle and colleagues (2002) and Stark and colleagues (2003) obtained no amygdala activation during the fear condition. This may suggest that the amygdala is not even necessary in fear processing. Recent studies by Tsuchiya, Moradi, Felsen, Yamazaki, and Adolphs (2009) and Piech and colleagues (2010) on SM found that the patients speed performance on a rapid detection task of fear-related stimuli was completely normal. This result, as the authors remark, suggests that the amygdala is not essential, at least for early stages of fear processing.

Besides these findings, several studies have found a broader role for the amygdala that involves processing both positive and negative emotion expressions. For instance, Breiter and colleagues (1996) found amygdala activation for happy versus neutral faces. In a study by Garavan, Pendergrass, Ross, Stein, and Risinger (2001), subjects viewed pictures that varied in emotional content (positive versus negative valence) while undergoing fMRI scanning. Amygdala activation was significantly increased for both positively and negatively valenced stimuli. In an fMRI study, Wright, Martis, Shin, Fischer, and Rauch (2002) evaluated human brain responses to simple drawings of emotional and neutral facial expressions. Significantly increased fMRI signals were found in the amygdala in response to angry and happy schematic faces. Other studies (Kim, Somerville, Johnstone, Alexander, & Whalen, 2003; Wang, McCarthy, Song, & LaBar, 2005) suggest that amygdala's role in emotion processing extends to facial expressions of surprise and sadness. Some studies have shown that the amygdala responds across all emotional expression conditions. Yang and colleagues (2002) examined amygdala responses to the perception of happy, angry, sad, and fearful facial expressions compared to neutral expressions. They found that all four facial expressions, including happy faces, were associated with reliable bilateral activation of the amygdala. Similarly, in a study by Winston, O'Doherty, and Dolan (2003), subjects viewed morphed emotional faces displaying low and

high intensities of disgust, fear, happiness, or sadness under two different task conditions. The amygdala responded to high-intensity expressions of all the four basic emotions, suggesting that the amygdala is involved in the perceptual processing of a range of emotions, rather than fear only. In an fMRI study (Fitzgerald, Angstadt, Jelsone, Nathan, & Phan, 2006), 20 subjects viewed photographs displaying fearful, disgusted, angry, sad, neutral, and happy facial expressions. The left amygdala was activated by each condition separately (emotional or non-emotional) and its response was not selective for any particular emotion category.

But what can be said about the early finding showing SM's poor performance on the fear recognition task? First, it is notable that although SM exhibited poor performance on recognizing fear from facial expressions, she could recognize fear from complex visual scenes and tone of voice. But apart from this, surprisingly, after more than a decade of study on SM, Adolphs and colleagues (2005) reported that SM's impaired recognition stems from her inability to make use of diagnostic information from the eye region that is normally essential for recognizing fear, and that this inability is related to her lack of spontaneous fixation on the eye region of faces. Interestingly, SM's recognition of fearful expressions became entirely normal when she was instructed to look at the eyes. It is also notable that discrimination of sadness and anger also makes substantial use of the eye region. This explains why impaired recognition of sadness and anger, along with fear, has been reported after amygdala damage. Finally, why did SM appear to be highly selective for fear recognition compared to other negative emotions? As Adolphs and colleagues (2005) remark, this is probably attributable to her ability to make compensatory use of information outside the eye region for those other emotions, a strategy insufficient with fear recognition.

Taken together, evidence from lesion and functional brain imaging studies challenge the notion that the amygdala is specialized specifically as a fear processor, and suggest a more general purpose role which encompasses processing multiple emotions and expressions of affect. Studies show that the amygdala is involved in processing stimuli that are ambiguous, unpredictable, and have particular behavioral salience (Adolphs, 2008, 2010; Whalen, 2007). Similarly, Sander and colleagues (2003) suggest that the amygdala is a system with a variety of cortical and subcortical projections that supply information about the properties of the stimulus as well as the ongoing needs of the organism. On this account, the amygdala is an evolved system that acts as a *relevance* detector, a broader functional category that is not restricted to fear processing or any specific emotion category.

Relatedly, Adolphs reminds us that the amygdala is a complex collection of 13 nuclei extensively connected with many other cortical and subcortical structures, and thus accounts of its function cannot do justice to its location unless they put it in a dense web of connections (2010, p. 42). The connections are so complex that researchers have even challenged the concept of "the amygdala" as a structural or functional unit (Swanson & Petrovich, 1998). Taking these points into account, the emerging picture is a conception of the amygdala that hardly meshes with Goldman's view of the amygdala as a distinct and specific fear processor.

0.4.3 Disgust and the Insula

Unlike the simulational evidence for fear, which was based only on results from lesion studies, the simulational argument for disgust relies on two sorts of evidence from both brain imaging and lesion studies, making disgust "an even clearer case than fear" (Goldman, 2006, p. 117). Goldman reports an fMRI study in which Phillips and colleagues (1997) obtained insula activation during observation of disgusted facial expressions such that, he remarks, adjacent regions such as the amygdala were not activated. In a different imaging study by Wicker and colleagues (2003), subjects inhaled odorants producing a strong feeling of disgust and observed video clips showing facial expression of disgust. The results by Wicker and colleagues show activation of the anterior insula and to a lesser extent activation of the anterior cingulate cortex in both conditions. Besides, evidence from lesion studies shows a pattern of a paired deficit in the experience and recognition of disgust. Calder, Keane, Manes, Antoun, and Young (2000) examined a patient NK with insula and basal ganglia damage. NK's score on a disgust experience questionnaire and performance in the disgust category of a facial emotion recognition task were lower than controls. Overall, the findings, according to Goldman's account, suggest that the same neural region, the insula, is specifically and selectively responsible for both experience and recognition of disgust. Therefore, disgust recognition is simulational (Goldman, 2006; Goldman & Sripada, 2005).

Two points to the contrary of the simulational interpretation of the Calder and colleagues study and the patient NK are in order. First, it is notable that although a deficit in disgust processing was observed in NK, the impairments involved processing of other emotions as well. During tests of non-verbal emotional sounds, NK showed a deficit in recognizing surprise and incorrectly labeled disgust as fear and anger, when interpreting JACFEE facial expressions, NK's recog-

nition of contempt was impaired, and he miscategorized disgust as anger and contempt, and for the Ekman and Friesen faces, NK incorrectly categorized disgust facial expression as anger. These are all ignored in the simulation argument. Could NK's impairments result from a more general deficit in visual processing? This might be the case especially because Phillips and colleagues (1998) didn't find insula or basal ganglia activation in response to presentation of *vocal* expressions of disgust. Besides, in emotion recognition tests using pictures of scenes (compared to recognition from faces), NK showed no difficulty at all in recognizing any emotion, including disgust. Given the total spectrum of findings from these studies, one can hardly think NK's impairments result from a deficit in a system specifically involved in disgust processing.

Second, other studies show preserved disgust experience and recognition in patients with damage similar to NK's. It is notable that patients similar to NK are rare, and for a decade NK was the only patient with focal brain damage to the basal ganglia and insular cortex. Recently, however, Straube and colleagues (2010) have reported a patient MK with a comparable lesion of the insula and basal ganglia. If the insula and basal ganglia are reliably involved in disgust processing, MK should show at least some deficits in the processing of disgust stimuli. To examine MK, Straube and colleagues used tests and methods similar to those used by Calder and colleagues (2000) to study NK. Contrary to the findings from Calder and colleagues, none of the tests by Straube and colleagues revealed a deficit in disgust processing for MK. Compared to healthy controls, MK showed no impairments in the recognition or experience of disgust, nor any notable impairment in the recognition and experience of other emotions. This finding is further confirmed by a more recent finding from a study by Couto and colleagues (2013) on a patient GG with focal insular damage. Like MK, GG showed no impairment in emotion recognition. The two studies show preserved disgust experience and recognition despite exclusive focal damage to the basal ganglia and insular cortex.

Because focal brain damage to the insula and basal ganglia is infrequent, other types of lesion studies might be illuminating here. Patients with Huntington's disease (HD), a neurodegenerative disorder that affects the basal ganglia and insula, are of primary interest in this area.[9] The primary evidence which inspired the idea that disgust

[9]Also, two types of psychiatric disorders, obsessive compulsive disorder and Tourette's syndrome, are informative in understanding the role of the basal ganglia and the insula in disgust recognition (see, for example, Calder, Lawrence, & Young, 2001). But here I confine my discussion only to results from studies on the Huntington's disease for the sake of brevity and greater relevance.

processing might be associated with a specific neural region came from neuropsychological findings in patients with Huntington's disease. HD patients are ignored in Goldman's account. In an early study, Sprengelmeyer and colleagues (1996) found HD patients with severe impairments in disgust processing during facial and vocal emotion recognition tests. But even in this early study, the average rate at which disgust was detected was below the next most badly affected emotion, fear, and in addition, the patients had severe problems in discriminating fear from anger. The study showed impairment in recognition of most emotions, with some emotions impaired more than others. In a follow-up study, Sprengelmeyer and colleagues (1997) further examined emotion recognition at the individual level in two HD patients, HL and UJ. The results showed severe impairments in the recognition of disgust and fear, but not disgust only, and one patient, UJ, had problems involving the misrecognition of fear as anger.

Several other studies confirm that HD impairments usually involve recognition of several emotions, rather than disgust only. Milders, Crawford, Lamb, and Simpson (2003) compared the performance of HD patients and gene-carriers of HD on two sets of emotion recognition tests, and found that HD patients were impaired at recognizing several expressions, including sadness, anger, disgust, and fear compared to healthy controls and asymptomatic gene-carriers. Interestingly, there was no indication of a selective impairment in disgust recognition. Further testing on selective impairment revealed that the patients were in fact significantly more impaired on other negative emotions, for instance, fear, than on disgust. The result was corroborated by three more emotion recognition studies, one on 475 HD patients, that found a decline in recognition of all negative emotions, including sadness, fear, disgust, anger, and surprise (Ille et al., 2011; Johnson et al., 2007; Snowden et al., 2008). Taken together, the data coming from emotion recognition in HD patients go against the simulational conception of a region as specifically a disgust processor. So much for the evidence from lesion studies.

Turning now to brain imaging studies, there is much evidence to suggest that disgust recognition is not, in fact, isolated to a specific region or network in the brain. First, that the insula is activated, even consistently, for disgust does not necessarily mean that the insula is specifically a disgust processor. Like what was said about the amygdala, to show that the insula is specifically a disgust processor, it must be demonstrated that the difference in activation obtained for disgust cannot be obtained for other emotion categories. However, the fMRI studies (Phillips et al., 1997; Wicker et al., 2003) on which Goldman's account relies focus exclusively on disgust; they compare

the insula activation for disgust relative to neutral faces only. But to find the desired specificity, the insula activation must be compared with respect to other emotions as well. Several studies, however, show that, upon such comparison, one can hardly conclude that these regions are specifically disgust processors. For instance, Schienle and colleagues (2002) presented subjects with pictures displaying a wide variety of different disgust and fear elicitors. The results show insula activation during the fear condition and amygdala activation during the disgust condition. The finding, as the authors remark, accords with the notion of the insula as a region involved in affective tasks without focusing on any specific emotion, a view which is also similar to Damasios (Damasio et al., 2000) conception of the insula as part of a central circuit concerned with monitoring emotional states in general.

In addition, Phillips and colleagues (1998) found activation in the insula and basal ganglia for fear, and in the amygdala in response to disgust.[10] In another study, Stark and colleagues (2003) found significant amygdala activation during the disgust condition. Similar brain structures were activated when Stark and colleagues contrasted the disgusting and fear-inducing pictures with the affectively neutral pictures. The results support the conception that fear and disgust are processed in similar brain structures.

Second, even the link between insula activation and disgust has not been entirely consistent. Phillips and colleagues (1998) didn't find insula or basal ganglia activation in response to auditory disgust stimuli. Schienle, Schfer, Stark, Walter, and Vaitl (2005b) analyzed data from 63 subjects across 4 studies to see if the insula, the amygdala, the orbitofrontal cortex, and the medial prefrontal cortex would be involved in disgust processing. Whereas subjects experienced intense disgust feelings (based on a self-report questionnaire), the study found no insular activation during the disgust conditions. This presumably undermines the notion of the insula as necessarily involved in disgust processing.

Third, several studies (Hutchison, Davis, Lozano, Tasker, & Dostrovsky, 1999; Jackson et al., 2005; Peyron, Laurent, & Garca-Larrea, 2000) show that the insula is involved in pain processing. For instance, Jackson and colleagues (2005) found that the insula significantly responds to the perception of painful situations in photographs.[11] The

[10]The number of activated voxels was greater in response to disgust and fear in the insula and the amygdala, respectively, but the point here is that the regions are not specifically disgust or fear processors.

[11]This and other studies show that several regions in addition to the insula are involved in pain processing, including the anterior cingulate, the cerebellum, and the thalamus.

findings from this and other studies hardly fit with the conception of the insula as specifically a disgust processor.

Finally, various findings demonstrate that brain regions other than the insula and basal ganglia are involved in disgust processing. Evidence from several studies (Gorno-Tempini et al., 2001; Schienle et al., 2005b, 2006; Winston et al., 2003) show amygdala activation in response to disgust-inducing stimuli. For instance, Schienle and colleagues (2006) compared neural responses of two types of disgust elicitors (pictures of contamination and humiliation) with fear-relevant and neutral scenes. The results show that both disgust conditions involved activation in the amygdala, the occipitotemporal cortex, and the orbitofrontal cortex, but no significant activation in the insula.[12]

Taken together, the above results from brain imaging and lesion studies speak against the simulational view that a distinct region is specifically a disgust processor. Instead, the data strongly suggests that several regions are involved in disgust processing, that insula is involved in the processing of several emotions, and that disgust and fear share components of an integrative and distribute system at least three of which are the insula, the basal ganglia, and the amygdala.

0.4.4 Anger and Dopamine Level

I now briefly discuss Goldman's argument for a third emotion, anger. Results from a study by Lawrence and Calder (2004) shows that dopamine level plays an important role in the experience of anger in rats and other species, and the study by Lawrence, Calder, McGowan, and Grasby (2002) suggests that administration of the dopamine antagonist sulpiride to healthy subjects selectively impairs facial recognition of anger. These findings, according to Goldman, demonstrate that the same substrate, dopamine level, is specifically involved in experience and recognition of anger. Therefore, anger recognition is simulational (Goldman, 2006; Goldman & Sripada, 2005).

I point out several problems with this argument. First, findings show that deficits in anger recognition occur independently of changes in dopamine levels. As discussed above, subjects with amygdala damage ((Adolphs et al., 1994, 1995; Blair et al., 1999; Scott et al., 1997; Sprengelmeyer et al., 1999), insula and basal ganglia damage ((Calder et al., 2000; Sprengelmeyer et al., 1996, 1997), and HD patients (Johnson et al., 2007; Milders et al., 2003; Snowden et al., 2008) show deficits in anger recognition without being subject to any specific dopamine

[12]The study by Schienle and colleagues (2006) was an attempt to replicate the data from a previous study by Wright, Shapira, Goodman, and Liu (2004), who were unable to image the amygdala with their 3T scanner.

level manipulation.

Second, the findings by Lawrence and colleagues (2002) are the result of a single study, not replicated by other studies. Interestingly, studies show that other pharmacological manipulations can also affect anger recognition. For instance, the administration of ethanol (Borrill, Rosen, & Summerfield, 1987) and diazepam (Blair & Curran, 1999) impairs anger recognition from facial expressions. Furthermore, dopamine is involved in other cognitive functions as well, for instance, its level modulates performance in working memory tasks (Durstewitz & Seamans, 2002). These findings hardly mesh with the simulational conception of dopamine level as specifically an anger processor.

Third, simulation is a process that duplicates or resembles another process (Goldman, 2006, p. 37). Dopamine, however, is a neurotransmitter which is obviously a *substance* and not a process. Of course, dopamine might be involved in processes that subserve anger recognition, but in this case, its role would be too general to be interesting for the purpose of Goldman's argument.

Better prospects might be found if an ST advocate looked for a region, instead of a substance, specifically involved in anger processing. So, like the simulation arguments for fear and disgust, one can find evidence that, for instance, links anger either to the orbitofrontal cortex (Blair et al., 1999) or the prefrontal cortex (Monk et al., 2006). However, from what has been discussed, one can see this position is hardly viable. It is not hard to find evidence showing that the orbitofrontal cortex, the prefrontal cortex, or any other possibly relevant region, is neither specifically involved in anger processing, nor that anger recognition relies solely on one of these possible brain structures. The pattern we found for fear and disgust holds not only for anger but for all other emotions and sensations as well; emotion and sensation recognition occurs in multiple integrative and distributed brain structures.

0.4.5 Conclusion and Simulationist Response

It is important to recap what we have done so far. The main purpose of the paper was to examine Goldman's simulational account of emotion recognition and his argument against information-rich explanations. On Goldman's account of emotion recognition, the brain during emotion recognition duplicates or resembles the brain during emotion experience such that for an emotion E_1, there is a region R_1 that instantiates both experience and recognition of E_1. Besides, deficits in R_1 would impair both experience and recognition of E_1. Thus, Goldman concludes, emotion recognition is simulational. How-

ever, I have shown that a closer examination of the evidence reveals that the brain during emotion recognition does not resemble the brain during emotion experience. In other words, the processes that underlie emotion recognition can be dissociated from the processes that underlie emotion experience. Several studies show that the brain regions Goldman considers as emotion-specific processors are often involved in the processing of several emotions and that the process of recognizing an emotion occurs in multiple brain structures. In addition, evidence demonstrates intact emotion experience and recognition (e.g., fear) despite severe deficit in the allegedly emotion-specific brain structure (e.g., the amygdala). Therefore, emotion recognition is not simulational, at least not in the sense intended by Goldman.

Further, Goldman argues that ST has more explanatory power because TT fails to explain selective impairments—that is, TT fails to explain why patients impaired in recognizing an emotion E_1 have no difficulty in recognizing other emotions E_2, E_3, and so on (Goldman, 2009b, pp. 145146). However, this argument hinges on the assumption that each of the basic emotions is specifically processed in a distinct brain structure. Careful examination of the evidence reveals that the reported brain regions are not emotion-specific processors. Indeed, that is why patients diagnosed with a deficit in recognition of an emotion are often impaired in recognition of several emotions. Thus, TT theorists need not be concerned about selective impairments.

To be clear, I am not denying the relative contribution or significance of any brain structure. Certainly, brain imaging studies help to spot regional changes and identify which brain area is most active for a certain emotion, but perhaps this identification is not the critical factor. Instead, as McIntosh (2004) remarks, the contribution of a brain region, or any neural element, to a mental function depends on that brain region's neural context; that is, it depends on the status of the other anatomically related regions at that point in time (McIntosh, 2004, p. 176). Under this understanding, even Brocas aphasia should not be attributed simplistically to Brocas area (Shimamura, 2010). Of course, results from brain imaging together with findings from lesion studies can provide convincing evidence that a region, for example, the amygdala, contributes to recognition of an emotion, for example, fear, but the crucial point is that this contribution ought not be interpreted in the context of brain region specification. Otherwise, we cannot explain why patients with lesions in specific brain regions rarely, if ever, have impairment in a single emotion, why a region active for an emotion is not active for every instance of that emotion category, and why specific brain regions respond to more than one emotion category.

Advocates of simulation theory would likely object that I have misdescribed the simulational account of emotion recognition. It might be objected that, while simulation theory is forced to the claim that the same brain region is used in both experience and recognition of an emotion, simulation is not committed to the thesis that each emotion is processed in a distinct brain structure. Why, the objection goes, cannot a simulationist think that the same brain region is responsible for different emotions, but just think that which emotions the brain area will "produce" depend on a particular pattern of activations. Or, that all emotions are produced by a unique and identical network of brain structures, but different emotions are simply a different pattern of activation in this network?

I think this objection is unsuccessful. First, if the same brain region R, or a network of brain structures, is responsible for different emotions, then damage to the region would impair recognition of *all* emotions. If so, the simulational account would fail to explain impairment in recognition of a specific emotion. But given the evidence we have examined, a reasonable way to avoid this problem is just to abandon the specific and selective impairment stuff and admit that emotion recognition occurs in a network of brain structures. However, giving up selective impairments does not seem to be an option for ST, as least not as far as Goldman's account is concerned. It is assumed that a process P qualifies as a simulation of P′ only if P duplicates or resembles (here in neural terms) P′ in significant respects (Goldman, 2006, p. 26). But how do the brain processes involved in emotion recognition duplicate or resemble the processes involved in emotion experience if emotion recognition is processed by a network of different brain structures? If recognition of an emotion E activates brain regions r_1, r_2, ..., r_n in a network of brain structures R, then recognition of E qualifies as simulational only if all (or a significant part of) the regions in R which are active when recognizing E also become active when experiencing E. However, this form of similarity has never been supported by simulation theorists nor to my knowledge by any independent empirical evidence. But it is notable that even if we suppose, for the sake of argument, that emotion experience and recognition occur in multiple, yet similar, brain structures, this form of similarity fits well with information-rich explanations of emotion recognition.

It seems we are faced with a dilemma. Either we abandon selective impairments and believe that emotion recognition occurs in multiple, but similar, brain structures, in which case we have no empirical evidence to provide backing for this possibility. Or, we stick to the notion of selective/specific impairment, in which case we will end up with a

simulational account which hardly squares with empirical evidence on how, in fact, emotion recognition is processed. Since Goldman's contributions, several authors have integrated various other contributions in the spirit of simulation theory. Does the evidence we have reviewed represent a challenge to these other simulational perspectives? The answer mainly depends on what we *mean* by simulation, but, while we should carefully distinguish between simulation and broader conceptual terms, such as "imitation" and "empathy"[13], it is clear that any simulational account which explains emotion recognition on the basis of specific, selective, and consistent processing in the brain will face similar problems.

[13]We must distinguish between simulation for emotion recognition which involves attributing an emotion to a target from empathy which is simply an emotional contagion or experience sharing. So, as Goldman remarks, whereas, for example, mirror processes might be involved in simulation as well as imitation or empathy, the role of mirror neurons in simulation does not necessarily follow from their role in imitation or empathy (Goldman, 2008, pp. 311–312).

0.4.6 Bibliography

Adolphs, R. (2008). Fear, faces, and the human amygdala. Current Opinion in Neurobiology, 18, 166-172.

Adolphs, R. (2010). What does the amygdala contribute to social cognition? Annals of the New York Academy of Sciences, 1191, 42-61.

Adolphs, R., Gosselin, F., Buchanan, T. W., Tranel, D., Schyns, P., & Damasio, A. R. (2005). A mechanism for impaired fear recognition after amygdala damage. Nature, 433, 68-72.

Adolphs, R., Damasio, H., Tranel, D., & Damasio, A. R. (1996). Cortical systems for the recognition of emotion in facial expressions. The Journal of Neuroscience, 16, 7678-7687.

Adolphs, R., & Tranel, D. (2000). Emotion recognition and the human amygdale. In J. P. Aggleton (Ed.), The amygdala: A functional analysis (2nd ed., pp. 587-630). Oxford: Oxford University Press.

Adolphs, R., Tranel, D., Damasio, H., & Damasio, A. (1994). Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. Nature, 372, 669-672.

Adolphs, R., Tranel, D., Damasio, H., & Damasio, A. R. (1995). Fear and the human amygdala. The Journal of Neuroscience, 15, 5879-5891.

Adolphs, R., Tranel, D., Hamann, S., Young, A. W., Calder, A. J., Phelps, E. A., ... Damasio, A. R. (1999). Recognition of facial emotion in nine individuals with bilateral amygdala damage. Neuropsychologia, 37, 1111-1117

Blair, R. J., & Curran, H. V. (1999). Selective impairment in the recognition of anger induced by diazepam. Psychopharmacology, 147, 335-338.

Blair, R. J. R., Morris, J. S., Frith, C. D., Perrett, D. I., & Dolan, R. J. (1999). Dissociable neural responses to facial expressions of sadness and anger. Brain, 122, 883-893.

Borrill, J. A., Rosen, B. K., & Summerfield, A. B. (1987). The influence of alcohol on judgement of facial expression of emotion. The British Journal of Medical Psychology, 60, 71-77.

Botterill, G., & Carruthers, P. (1999). The philosophy of psychology. New York, NY: Cambridge University Press.

Breiter, H. C., Etcoff, N. L., Whalen, P. J., Kennedy, W. A., Rauch,

S. L., Buckner, R. L., ... Rosen, B. R. (1996). Response and habituation of the human amygdala during visual processing of facial expression. Neuron, 17, 875-887.

Calder, A. J., Lawrence, A. D., & Young, A. W. (2001). Neuropsychology of fear and loathing. Nature Reviews Neuroscience, 2, 352-363.

Calder, A. J., Keane, J., Manes, F., Antoun, N., & Young, A. W. (2000). Impaired recognition and experience of disgust following brain injury. Nature Neuroscience, 3, 1077-1078.

Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. The Journal of Philosophy, 78, 67-90.

Couto, B., Sedeo, L., Sposato, L. A., Sigman, M., Riccio, P. M., Salles, A., ... Ibanez, A. (2013). Insular networks for emotional processing and social cognition: Comparison of two case reports with either cortical or subcortical involvement. Cortex; a Journal Devoted to the Study of the Nervous System and Behavior, 49, 1420-1434.

Damasio, A. R., Grabowski, T. J., Bechara, A., Damasio, H., Ponto, L. L. B., Parvizi, J., & Hichwa, R. D. (2000). Subcortical and cortical brain activity during the feeling of self-generated emotions. Nature Neuroscience, 3, 1049-1056.

Decety, J., & Jackson, P. L. (2004). The functional architecture of human empathy. Behavioral and Cognitive Neuroscience Reviews, 3, 71-100.

Dinstein, I., Thomas, C., Behrmann, M., & Heeger, D. J. (2008). A mirror up to nature. Current Biology, 18, R13-R18.

Durstewitz, D., & Seamans, J. K. (2002). The computational role of dopamine D1 receptors in working memory. Neural Networks, 15, 561-572.

Fitzgerald, D. A., Angstadt, M., Jelsone, L. M., Nathan, P. J., & Phan, K. L. (2006). Beyond threat: Amygdala reactivity across multiple expressions of facial affect. NeuroImage, 30, 1441-1448.

Gallese, V. (2001). The "shared manifold" hypothesis: From mirror neurons to empathy. Journal of Consciousness Studies, 8, 33-50.

Gallese, V., & Goldman, A. I. (1998). Mirror neurons and the simulation theory of mindreading. Trends in Cognitive Sciences, 2, 493-501.

Garavan, H., Pendergrass, J. C., Ross, T. J., Stein, E. A., & Risinger,

R. C. (2001). Amygdala response to both positively and negatively valenced stimuli. Neuroreport, 12, 2779-2783.

Godfrey-Smith, P. (2005). Folk psychology as a model. Philosophers' Imprint, 5(6), 1-16.

Goldman, A. I. (2006). Simulating minds. New York, NY: Oxford University Press.

Goldman, A. I. (2008). Mirroring, mindreading, and simulation. In J. A. Pineda (Ed.), Mirror neuron systems (pp. 311-330). New York, NY: Humana Press.

Goldman, A. I. (2009a). Mirroring, simulating and mindreading. Mind & Language, 24, 235-252.

Goldman, A. I. (2009b). Simulation theory and cognitive neuroscience. In D. Murphy & M. Bishop (Eds.), Stich and his critics (pp. 137-166). Malden, MA: Wiley.

Goldman, A. I. (2012). Theory of mind. In E. Margolis & S. Stich (Eds.), The Oxford handbook of philosophy of cognitive science (pp. 402-424). New York, NY: Oxford University Press.

Goldman, A. I., & Jordan, L. (2013). Mindreading by simulation: The roles of imagination and mirroring. In S. Baron-Cohen, M. Lombardo, & H. Tager-Flusberg (Eds.), Understanding other minds: Perspectives from developmental social neuroscience (pp. 448-466). New York, NY: Oxford University Press.

Goldman, A. I., & Sripada, C. S. (2005). Simulationist models of face-based emotion recognition. Cognition, 94, 193-213.

Gordon, R. M. (1986). Folk psychology as simulation. Mind & Language, 1, 158-171.

Gorno-Tempini, M. L., Pradelli, S., Serafini, M., Pagnoni, G., Baraldi, P., Porro, C., ... Nichelli, P. (2001). Explicit and incidental facial expression processing: An fMRI study. NeuroImage, 14, 465-473.

Hamann, S. B., Stefanacci, L., Squire, L. R., Adolphs, R., Tranel, D., Damasio, H., & Damasio, A. (1996). Recognizing facial emotion. Nature, 379, 497.

Heal, J. (1986). Replication and functionalism. In J. Butterfield (Ed.), Language, mind, and logic (Vol. 14, pp. 135-150). Cambridge: Cambridge University Press.

Heal, J. (1995). How to think about thinking. In M. Davies & T. Stone (Eds.), Mental Simulation (pp. 33-52). Hoboken, NJ: Blackwell.

Hickok, G. (2009). Eight problems for the mirror neuron theory of action understanding in monkeys and humans. Journal of Cognitive Neuroscience, 21, 1229-1243.

Hickok, G. (2014). The myth of mirror neurons: The real neuroscience of communication and cognition. New York, NY: Norton.

Hutchison, W. D., Davis, K. D., Lozano, A. M., Tasker, R. R., & Dostrovsky, J. O. (1999). Painrelated neurons in the human cingulate cortex. Nature Neuroscience, 2, 403-405.

Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., & Rizzolatti, G. (1999). Cortical mechanisms of human imitation. Science, 286, 2526–2528.

Ille, R., Holl, A. K., Kapfhammer, H.-P., Reisinger, K., Schfer, A., & Schienle, A. (2011). Emotion recognition and experience in Huntingtons disease: Is there a differential impairment? Psychiatry Research, 188, 377-382.

Jackson, P. L., Meltzoff, A. N., & Decety, J. (2005). How do we perceive the pain of others? A window into the neural processes involved in empathy. NeuroImage, 24, 771-779.

Johnson, S. A., Stout, J. C., Solomon, A. C., Langbehn, D. R., Aylward, E. H., Cruce, C. B., ... Group, the P.-H. I. of the H. S. (2007). Beyond disgust: Impaired recognition of negative emotions prior to diagnosis in Huntington's disease. Brain, 130, 1732-1744.

Keysers, C., Wicker, B., Gazzola, V., Anton, J.-L., Fogassi, L., & Gallese, V. (2004). A touching sight: SII/PV activation during the observation and experience of touch. Neuron, 42, 335-346.

Kim, H., Somerville, L. H., Johnstone, T., Alexander, A. L., & Whalen, P. J. (2003). Inverse amygdala and medial prefrontal cortex responses to surprised faces. Neuroreport, 14, 2317-2322.

Lawrence, A. D., & Calder, A. J. (2004). Homologizing human emotions. In D. Evans & P. Cruse (Eds.), Emotion, evolution, and rationality (pp. 1548). New York, NY: Oxford University Press.

Lawrence, A. D., Calder, A. J., McGowan, S. W., & Grasby, P. M. (2002). Selective disruption of the recognition of facial expressions of anger. Neuroreport, 13, 881-884.

Maibom, H. (2009). In defence of (model) theory theory. Journal of Consciousness Studies, 16, 360-378.

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

McIntosh, A. R. (2004). Contexts and catalysts: A resolution of the localization and integration of function in the brain. Neuroinformatics, 2, 175-182.

Milders, M., Crawford, J. R., Lamb, A., & Simpson, S. A. (2003). Differential deficits in expression recognition in gene-carriers and patients with Huntingtons disease. Neuropsychologia, 41, 1484-1492.

Monk, C. S., Nelson, E. E., McClure, E. B., Mogg, K., Bradley, B. P., Leibenluft, E., ... Pine, D. S. (2006). Ventrolateral prefrontal cortex activation and attentional bias in response to angry faces in adolescents with generalized anxiety disorder. American Journal of Psychiatry, 163, 1091-1097.

Nichols, S., & Stich, S. P. (2003). Mindreading. An integrated account of pretence, self-awareness, and understanding other minds. New York, NY: Oxford University Press.

Pellegrino, G. di, Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: A neurophysiological study. Experimental Brain Research, 91, 176-180.

Perner, J., & Kuhberger, A. (2005). Mental simulation: Royal road to other minds? In B. F. Malle & S. D. Hodges (Eds.), Other minds: How humans bridge the divide between self and others (pp. 166–181). New York, NY: Guilford Press.

Peyron, R., Laurent, B., & Garca-Larrea, L. (2000). Functional imaging of brain responses to pain. A review and meta-analysis. Neurophysiologie Clinique/Clinical Neurophysiology, 30, 263-288.

Piech, R. M., McHugo, M., Smith, S. D., Dukic, M. S., Meer, J. V. D., Abou-Khalil, B., & Zald, D. H. (2010). Fear-enhanced visual search persists after amygdala lesions. Neuropsychologia, 48, 3430-3435.

Phillips, M. L., Young, A. W., Scott, S. K., Calder, A. J., Andrew, C., Giampietro, V., ... Gray, J. A. (1998). Neural responses to facial and vocal expressions of fear and disgust. Proceedings of the Royal Society of London B: Biological Sciences, 265, 1809-1817.

Phillips, M. L., Young, A. W., Senior, C., Brammer, M., Andrew, C., Calder, A. J., ... David, A. S. (1997). A specific neural substrate for perceiving facial expressions of disgust. Nature, 389, 495-498.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. Cognitive Brain

Research, 3, 131-141.

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. Nature Reviews Neuroscience, 2, 661-670.

Sander, D., Grafman, J., & Zalla, T. (2003). The human amygdala: An evolved system for relevance detection. Reviews in the Neurosciences, 14, 303–316.

Schienle, A., Schfer, A., Stark, R., Walter, B., & Vaitl, D. (2005). Neural responses of OCD patients towards disorder-relevant, generally disgust-inducing and fear-inducing pictures. International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology, 57, 69-77.

Schienle, A., Schfer, A., Stark, R., Walter, B., & Vaitl, D. (2005b). Relationship between disgust sensitivity, trait anxiety and brain activity during disgust induction. Neuropsychobiology, 51, 86-92.

Schienle, A., Schfer, A., Hermann, A., Walter, B., Stark, R., & Vaitl, D. (2006). fMRI responses to pictures of mutilation and contamination. Neuroscience Letters, 393, 174-178.

Schienle, A., Stark, R., Walter, B., Blecker, C., Ott, U., Kirsch, P., ... Vaitl, D. (2002). The insula is not specifically involved in disgust processing: An fMRI study. Neuroreport, 13, 2023-2026.

Scott, S. K., Young, A. W., Calder, A. J., Hellawell, D. J., Aggleton, J. P., & Johnsons, M. (1997). Impaired auditory recognition of fear and anger following bilateral amygdala lesions. Nature, 385, 254-257.

Shimamura, A. P. (2010). Bridging psychological and biological science: The good, bad, and ugly. Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 5, 772-775.

Siegle, G. J., Steinhauer, S. R., Thase, M. E., Stenger, V. A., & Carter, C. S. (2002). Cant shake that feeling: Event-related fMRI assessment of sustained amygdala activity in response to emotional information in depressed individuals. Biological Psychiatry, 51, 693-707.

Singer, T., Seymour, B., ODoherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. Science, 303, 1157-1162.

Snowden, J. S., Austin, N. A., Sembi, S., Thompson, J. C., Craufurd, D., & Neary, D. (2008). Emotion recognition in Huntingtons dis-

ease and frontotemporal dementia. Neuropsychologia, 46, 2638-2649.

Sprengelmeyer, R., Young, A. W., Schroeder, U., Grossenbacher, P. G., Federlein, J., Buttner, T., & Przuntek, H. (1999). Knowing no fear. Proceedings of the Royal Society of London B: Biological Sciences, 266, 2451-2456.

Sprengelmeyer, R., Young, A. W., Calder, A. J., Karnat, A., Lange, H., Hmberg, V., Rowland, D. (1996). Loss of disgust. Perception of faces and emotions in Huntingtons disease. Brain: A Journal of Neurology, 119, 1647-1665.

Sprengelmeyer, R., Young, A. W., Sprengelmeyer, A., Calder, A. J., Rowland, D., Perrett, D., Lange, H. (1997). Recognition of facial expressions: Selective impairment of specific emotions in Huntingtons disease. Cognitive Neuropsychology, 14, 839-879.

Stark, R., Schienle, A., Walter, B., Kirsch, P., Sammer, G., Ott, U., ... Vaitl, D. (2003). Hemodynamic responses to fear and disgust-inducing pictures: An fMRI study. International journal of psychophysiology: Official journal of the international organization of psychophysiology, 50, 225-234.

Stich, S. P., & Nichols, S. (1992). Folk psychology: Simulation or tacit theory? Mind & Language, 7, 35-71.

Straube, T., Weisbrod, A., Schmidt, S., Raschdorf, C., Preul, C., Mentzel, H.-J., & Miltner, W. H. R. (2010). No impairment of recognition and experience of disgust in a patient with a right-hemispheric lesion of the insula and basal ganglia. Neuropsychologia, 48, 1735-1741.

Swanson, L. W., & Petrovich, G. D. (1998). What is the amygdala? Trends in Neurosciences, 21, 323-331.

Tsuchiya, N., Moradi, F., Felsen, C., Yamazaki, M., & Adolphs, R. (2009). Intact rapid detection of fearful faces in the absence of the amygdala. Nature Neuroscience, 12, 1224-1225.

Wang, L., McCarthy, G., Song, A. W., & LaBar, K. S. (2005). Amygdala activation to sad pictures during high-field (4 Tesla) functional magnetic resonance imaging. Emotion, 5, 12-22.

Whalen, P. J. (2007). The uncertainty of it all. Trends in Cognitive Sciences, 11, 499-500.

Whalen, P. J., Shin, L. M., McInerney, S. C., Fischer, H., Wright, C. I., & Rauch, S. L. (2001). A functional MRI study of human amygdala responses to facial expressions of fear versus anger. Emotion,

1, 70-83.

Wicker, B., Keysers, C., Plailly, J., Royet, J.-P., Gallese, V., & Rizzolatti, G. (2003). Both of us disgusted in my insula: The common neural basis of seeing and feeling disgust. Neuron, 40, 655-664.

Winston, J. S., ODoherty, J., & Dolan, R. J. (2003). Common and distinct neural responses during direct and incidental processing of multiple facial emotions. NeuroImage, 20, 84-97.

Wright, P., He, G., Shapira, N. A., Goodman, W. K., & Liu, Y. (2004). Disgust and the insula: fMRI responses to pictures of mutilation and contamination. Neuroreport, 15, 2347-2351.

Wright, C. I., Martis, B., Shin, L. M., Fischer, H., & Rauch, S. L. (2002). Enhanced amygdala responses to emotional versus neutral schematic facial expressions. Neuroreport, 13, 785-790.

Yang, T. T., Menon, V., Eliez, S., Blasey, C., White, C. D., Reid, A. J., ... Reiss, A. L. (2002). Amygdalar activation associated with positive and negative facial expressions. Neuroreport, 13, 1737-1741.

0.5 Mindreading, Simulation, and Pragmatic Interpretation

Abstract

Following Grice (1989), pragmatic interpretation is often seen as an exercise in mindreading. Mindreading itself, however, has been understood in rather different ways. What cognitive mechanisms underpin the mindreading exhibited in utterance interpretation? According to one hypothesis, the simulation theory, understanding an utterance is achieved by imaginative projection: asking what would I have meant by that utterance if I were in the speakers situation. In the first part of this paper, I discuss several problems with this view, most importantly I argue that the simulation strategy is not only cognitively too demanding but virtually ineffective in utterance interpretation. Next, drawing on empirical evidence from three clinical populations, I show that, contrary to what simulation hypothesis predicts, deficits in pragmatic interpretation are not associated with simulation impairments, hence simulation cannot play any significant role in utterance comprehension.

0.5.1 Introduction

The ability to impute mental states to oneself and others, often called the theory of mind or mindreading, is fundamental to our social interaction. However, questions concerning the cognitive basis of this function continue to be the subject of sustained debate in the literature (Apperly, 2010; Botterill & Carruthers, 1999; Davies & Stone, 1995a, 1995b; Gallese, 2001; Gallese & Goldman, 1998; Goldman, 2006, 2008; Nichols & Stich, 2003; Josef Perner & Khberger, 2005; Saxe & Baron-Cohen, 2007; Stich, 2009). Whereas the early debates on mindreading were set up as a two-sided battle between the two dominant accounts of Theory- Theory and Simulation Theory, more recently theorists have more or less realized that mindreading is a complex phenomenon which, depending on the domain of application, can be underpinned by theory or simulation. Given this general consensus on the complexity of the phenomenon, the question concerning the underlying mechanisms of mindreading must be addressed with respect to the role mindreading ability plays in particular domains and in connection with different cognitive functions.

Of special interest in this respect is the domain of human communication, in particular, evidence from communication impairments can be illuminating here. For instance, individuals with autism spectrum disorder are impaired in verbal and non-verbal communication. Subjects with autism find their social environment incomprehensible and often treat people and objects alike (Baron-Cohen, Leslie, & Frith, 1985, p. 38). But autistic subjects known to be deficient in communication are also deficient in mindreading. Could this co-occurrence help in providing a better understanding of communication? In clinical pragmatics, explanation of communication disorders in terms of mindreading deficits has been more dominant than other explanations, for instance, explanations in terms of inferential load and relevance processing (Cummings, 2014).

This approach, however, requires we know what kind of mechanisms underlie mindreading. Depending on how we understand mindreading, different implications it has with respect to our understanding of pragmatic impairments. Consider, for instance, the study by Baron-Cohen et al. (1985) in which they administered a version of the Wimmer & Perner's (1983) false-belief test (the acid test for the presence of mindreading) to three groups of children: a group of 20 autistic children with relatively high mean IQ of 82, a group of 14 children with Downs syndrome with an average IQ of 64, and a group of 27 normal preschool children. Results for Downs syndrome and normal children were similar: 23 out of 27 (85%) normal subjects and 12 out of 14 (86%) Down syndrome children passed the test. By contrast, 16 out of 20 (80%) autistic subjects failed the test. Because autistic subjects had a relatively high IQ, and mentally retarded non-autistic subjects, subjects with Downs syndrome, are in general socially competent, one very plausible explanation is to understand autistic impairments in terms of a difficulty in mindreading, rather than to attribute the impairments to general intelligence. This explanation, of course, ultimately depends on how we understand mindreading. However, a clear understanding of mindreading has proved to be hard to come by. In the same study, Baron-Cohen proposed that autistic subjects are impaired in mindreading because the subjects fail to employ a *theory* of mind. Similarly, Leslie (Leslie, 1987; Leslie & Frith, 1988) argued that autistic subjects' impairments stem from deficits in metarepresentation. However, Gordon (1986) and Goldman (1989), two of the advocates of the simulation hypothesis, argued that the communication impairments in autistic subjects of Baron-Cohen study result from imagination and perspective taking impairments.

Yet there is a trade-off here because, while an account of mindreading helps to explain communication impairments, results from studies

on communication deficits can provide insight into a better understanding of the underlying mechanisms of mindreading. The aim of the present paper is to use empirical evidence from clinical pragmatics to see if the simulation hypothesis, one of the currently dominant accounts of mindreading, can account for the mindreading exhibited in communication, in particular, to evaluate the application of the simulation hypothesis to the domain of utterance interpretation.

0.5.2 Simulation and Utterance Interpretation

Following Grice (Grice, 1957, 1969, 1975, 1989), there is a general consensus that our communication does not consist of a sequence of disconnected remarks, but involves, in addition to linguistically decoding (in verbal communication), the ability of metarepresentation and expression and recognition of intentions. Non-sentential items such as flag signals, gestures, facial expressions, and physical postures have no syntactic structure or component that contribute to the meaning of the whole. How does a hand wave or a blue flag in yacht racing mean anything? An expression such as hand wave has a particular meaning in a communicative context because it is an expression of an intention mutually recognized by communicators. Even in a verbal communication, what is meant normally deviates from what is said, or from conventionally encoded information. The gap between what is said (linguistically coded information) and what is communicated cannot be bridged unless communicators engage in a cooperative effort in which they express and recognize each others mental states. To see this, consider Mr. A who is a philosopher and is writing a testimonial about a student. He writes: "Mr. X's command of English is excellent, and his attendance at tutorials has been regular". Of course, Mr. A is implicitly conveying that X is no good at philosophy, but the reader can recover the implicit meaning only to the extent that he is able to make certain assumptions about Mr. As mental states. For instance, that Mr. A knows that the reader expects more information than this. The reader knows that Mr. A could not be unable to provide more information because X is his student. Also, Mr. A knows that more relevant information than this is required, and also knows that the reader knows this. The assumptions guide the reader to notice that Mr. A is reluctant to give more information about X, which implicates that X is no good at philosophy.

Not only recognition of conversational implicatures but also identification of explicit content involves metarepresentational competence. In order to establish what a speaker intends to assert, a hearer must be able to, for instance, disambiguate and assign a reference, fix the scope

of quantifiers, resolve the interpretation of vague expressions, and resolve illocutionary indeterminacies (Sperber & Wilson, 2002; Wilson, 2005). Whether it is referential ambivalences to resolve, implicatures to identify, illocutionary indeterminacies to resolve, or metaphors, ironies or non-sentential expressions to interpret, people in a communicative context are deeply involved in spontaneous metarepresentation, either for what they trying to express or in recognizing what is expressed. But how is the mindreading involved in communication guided and accomplished? Consider the general pattern of metarepresentation in the Gricean schema for identification of conversational implicatures:

> He said that p; ... he could not be doing this unless he thought that q; he knows (and knows that I know that he knows) that I can see that the supposition that he thinks that q is required; he has done nothing to stop me thinking that q; he intend me to think, or at least willing to allow me to think, that q; and so he has implicated that q (Grice, 1989, p.31).

The working-out scheme can be described adequately as an exercise in reflective reasoning in which expression and recognition of communicative intentions involve the application of inferential abilities to rules, generalizations, or concepts of a theory of mind. Details on the nature of these generalizations and the theory are not important here. What matters is that according to this view utterance interpretation is an inferential and thoroughly metarepresentational process in which either a general theory of mind is applied to the domain of communication or alternatively a specialized theory of mind module is dedicated for use in the domain of communication.

The above approach to mindreading, often under the label of theory-theory, has been the dominant explanatory strategy in understanding a wide range of cognitive functions, including among the first the ability to comprehend linguistic behavior (Stich & Nichols, 1992). However, there is an alternative approach to mindreading, called the simulation theory, according to which we understand others by using our own mind as a model for others' mental life. The basic idea of the simulation hypothesis is well expressed in a study by Kahneman and Tversky (1981) in which participants were asked to consider two travelers, Mr. Crane and Mr. Tees, who were scheduled to leave the airport on different flights, at the same time. They traveled from town in the same limousine, which was caught in a traffic jam, and arrived at the airport 30 minutes late. Mr. Crane is told his flight left on time. Mr. Tees is told that his flight was delayed, and just left 5 minutes ago.

Who do you think is more upset? 96% of subjects said that Mr. Tees was more upset. How did they come up with this answer? On a simulationist account, "each subject would have put himself in each of the imaginary travelers shoes and have imagined he would have felt in that place" (Gallese & Goldman, 1998, p. 496). Likewise, Gordon (1986) argues that in understanding and predicting behavior, we simulate by answering the question "what would I do in that person's situation?". We, Gordon remarks, imaginatively project into the other's situation in the same way chess players do when playing against an opponent. Chess players, while "transformed in imagination", visualize the board from the other side and act accordingly.

While the above cases express the core idea of the simulation hypothesis, simulation theorists have provided more details on how simulational processes are executed. On a typical understanding of the simulation hypothesis, simulation is a process in which we understand others by imaginatively replicating a target's mental states, that is, by imaginatively generating states that stand as representational surrogates for those of the target. The imaginatively generated states, or pretend states, are then fed into mindreader's own mental mechanisms, for example into the decision-making system which momentarily operates 'off-line'(i.e. disengaged from its standard operations). In the final stage, an output is generated and attributed to the target. In Gordon's words, during simulation,

> Our decision-making or practical reasoning system gets partially disengaged from its 'natural'inputs and fed instead with suppositions and images (or their 'subpersonal'or 'sub-doxastic'counterparts). Given these artificial pretend inputs the system then makes up its mind what to do. Since the system is being run off-line, as it were, disengaged also from its natural output systems, its 'decision'isn't actually executed but rather ends up as an anticipation... of the other's behaviour. (Gordon, 1986, p.170)

This simulational strategy has been used to account for mindreading in different domains, including decision prediction (Goldman, 1989; Gordon, 1986), figuring out solutions to arithmetic questions (Heal, 1995), inference prediction (Stich & Nichols, 1995), and predicting the grammaticality judgments (Harris, 1992). How about comprehending linguistic behavior in intentional terms? Here again, simulation theorists deny any substantial role for a theory or such internally represented knowledge structure in understanding linguistic behavior. Goldman (1989) argues:

> Verbal communicators commonly make assumptions... My question is: how does a communicator proceed to estimate what pieces of information will be marshaled, or made salient, in the mind of the audience... The speaker cannot appeal to any such *theoretical* knowledge to make predictions of what is likely to be derived or calculated by the hearer. Nonetheless, speakers are evidently pretty good at making such predictions, more precisely, at predicting what kinds of 'implicatures', will be appreciated by an audience. How do they do that? Again, I suggest, by simulation. (Goldman, 1989, p.171-2)

Similarly, Currie & Ravenscroft maintain:

> Understanding a speaker's meaning where it differs from the meaning of what is said is a plausible candidate for something one would do by imaginative projection... Putting yourself in the speaker's shoes and asking 'What would I have meant by that?' would be a good way to solve the problem. (Currie & Ravenscroft, 2002)

But how the simulation strategy can explain the mindreading exhibited in communication, in particular in utterance interpretation? One difficulty with this suggestion is that the simulation hypothesis has been often used to account for mindreading in *predictive* cases: action prediction, decision prediction, inference prediction, grammaticality judgment prediction etc. Predictions, however, proceed by moving forward. For example, in action prediction, one proceeds forward from imaginatively generated (pretend) mental states, runs the states through his own decision-making system, and predicts the possible effect (an action) of those mental states. But an utterance is a generated piece of action that requires *explanation*. Unlike predictions, explanations are achieved by moving backward from an observed action to the mental states that have resulted in that action. Can simulation run backward? If it does, does it work in utterance interpretation too? It is argued that simulation can be involved in both predictive and retrodictive cases. How does retrodictive simulation work? Retrodictive, or backward, simulation requires figuring out what mental states did the target have that led him to that action. To achieve this, he simulator imagines himself in the target's situation and conjectures possible mental states that could have caused the observed action. Retrodictive simulation was first proposed in motor domain (understanding observable motor movements) by Gallese & Goldman (1998) and others (Fogassi et al., 2005; Iacoboni et al., 2005). But how, exactly, does the process could work in utterance interpretation? A

more elaborate account of backward simulation has been offered by Goldman (Goldman, 2006; Goldman & Sripada, 2005) in his model of "generate and test" strategy:

> The "generate" stage produces hypothesized states or state combinations that might be responsible for the observed (or inferred) evidence.... The "test" stage consists of trying out one or more of the hypothesized state combinations to see if it would yield the observed evidence.... One E-imagines being in the hypothesized combination of states, lets an appropriate mechanism operate on them, and sees whether the generated upshot matches the observed upshot. (Goldman, 2006, p.184)

On the generate and test model, to understand an action we generate one or more hypotheses about the mental states that might be responsible for that action. To test the hypotheses, we imaginatively pretend to be in the hypothesized mental states, then run them one at a time through our own decision-making system to see which one results in an output (action) that matches the observed action.

Several problems with the generate and test model arise immediately. To begin with, the generate-and-test strategy is not a purely simulational process. The generation of hypothesized mental states, as the possible cause of actions, relies on generalizations (a theory) about connections between particular actions and certain mental states that cause them. Without such generalizations, the 'generate' stage would never get off the ground. Suppose, as Goldman has noted, that the 'generate and test' strategy is a hybrid account that consist of a theory-driven stage (hypothesis generation) followed by a purely simulational mechanism (testing stage). Even so, there remains several difficulties to worry about, perhaps the most obvious being that understanding in terms of the generate and test strategy results in duplication of the target behavior. According to the generate and test strategy, understanding an action requires generating hypotheses that in order to be tested a replica of the target action must be generated: the mindreader's cognitive mechanism is provided with inputs, and next the output of the system is tested against the target action. If the generated output matches the target action, the hypothesis is verified and so simulator can make sense of the target action. However, whereas the generate and test might be the mechanism that operates in understanding certain emotions, for example, when observation of a disgusted facial expression results in similar feeling or facial expression in the observers (Wicker et al., 2003), this is not what happens when we understand others' actions: we do not understand actions by replicating observed actions.

Second, and relatedly, the generate and test strategy runs counter to the standard understanding of the simulation hypothesis. On the standard simulation, the practical reasoning system is taken 'off-line'—that is, momentarily disengaged from its natural inputs and outputs—and the inputs are imaginative and 'tagged' as belonging to the target. That is why simulation, under the standard understanding, never results in duplication or execution of simulated action, as it normally does in its usual operation, but instead returns an output (mental states) that is attributed to the target (for an overview of the simulation hypothesis, see Goldman, 1989, 2006; Gordon, 1986; Nichols & Stich, 2003). Contrary to the standard understanding, hypothesized mental states in the generate and test strategy result in generating an action which is then compared to the target action. The problem is that, either the simulating system runs off-line, in which case the generate and test strategy fails to accomplish action understanding because no action is produced that can be tested to understand the target action. Or, simulation operates on genuine inputs and returns an action output, but this results in duplicating actions and this is not what actually happens when we understand actions.

There are also other problems with the simulation hypothesis. The generate and test strategy is not only cognitively too demanding and impractical (generating and testing actions every time we make sense of an action), it is also virtually ineffective in communication. As discussed above, utterances and non-verbal behavior often greatly underdetermine what is communicated. To establish the speaker's meaning, there are indeterminacies that must be observed, including recognition of implicatures and other disambiguates. How does the generate and test strategy can bridge the gap between what is said and what is communicated? Perhaps simulation could be the possible strategy if there were one-to-one relations between actions and mental states, such that for any action there was a specific set of mental states that could give rise to that action. This clearly is not the case because the same action can be motivated by different intentions and there may be more than one interpretation for the same action. Besides, listing all possible interpretations does not seem to be an option because then again there must be a highly *theoretical* mechanism that selects, from a range of possible interpretations, the most relevant interpretation, the one that can be tested (in the test stage of simulation) and sounds as the best explanation of the target action.

An advocate of simulation might argue that cognitive load of the testing stage is facilitated by the range of possible actions constrained by previous experience and practicalities. However, that does not sound too promising because, while ordinary actions might be a re-

iteration of previous actions, many linguistic actions are wholly new. As Sperber and Wilson (2002) remark, the prior probability of most utterances ever occurring is close to zero. Thus, whereas the semantic complexity of ordinary intentions is limited by the relatively limited range of possible actions, there are no such limitations on the semantic complexity of speakers intentions (Sperber & Wilson, 2002, p. 17).

Still, some theorists might object that understanding simulation in terms of 'the generate and test strategy' or 'off-line simulation' is too narrow and constraining. Besides, not all simulation theorists argue for off-line simulation. Harris (1992) and Heal (1996) defend a version of simulation in which nothing corresponds to the off-line operation. Similarly, Currie & Ravenscroft (2002), while keeping the notion of imaginative inputs, have abandoned the idea of off-line operation, arguing that the whole notion of bringing the system off-line is problematic because it is an obvious fact that people are able to do mind-reading tasks while performing other actions (Currie & Ravenscroft, 2002, p. 70). It might be argued that, while simulation in its narrower sense faces certain difficulties, simulation under a broader interpretation is the strategy we adopt in utterance interpretation. So, whereas hypothesis generation and hypothesis testing might be somehow involved in backward simulation, neither are crucial in utterance interpretation because, the argument goes, simulation for mindreading is a species of mental simulation in general (Goldman, 2006) and this general sense of mental simulation essentially consists in the ability of perspective taking, the ability of placing oneself in imagination in the target's situation and (retrodictively) see what that action mean.

This broader sense of simulation is close to Gordon (1986), Harris (1992), and Heal's (1996) conception of simulation. This conception of simulation is also advocated by other simulation theorists. For instance, Currie & Ravenscroft (2202) have argued that during simulation it is only the perspective taking part that is, properly speaking, simulation (Currie & Ravenscroft, 2002, p.54). Also, Goldman's recent work on high-level mindreading characterizes simulation in terms of the general ability of enactment imagination, visual perspective taking, visualization and motor imagery (Goldman, 2006, 2008, 2012; Goldman & Jordan, 2013). This is nicely illustrated in an example by Goldman:

> I am planning tonight's dinner. I just purchased a white beans and artichoke salad, which might nicely combine with a bed of leafy greens already in the refrigerator. To test the appeal of this combination, I visualize the white beans and pale green artichoke hearts against the background of the dark green (and red) leafy ingredients. This

> act of visual imagination is an instance of E-imagination. (Goldman, 2006, p. 149)

Visual and imaginative perspective taking, as Baron-Cohen remark, requires primary (first-order) representations only and "can be performed using the strategy of mental rotation on primary representations" (Baron-Cohen, 1988, p. 394). If simulation in this broader sense is the mechanism that is responsible for mindreading in utterance interpretation, then deficits in perspective taking must be associated with pragmatic impairments. In the next section of the paper, I argue that empirical evidence from different clinical populations speak against this association.

0.5.3 Dissociation Between Pragmatic impairments and Simulation Deficits

Communication impairments are among the key symptoms of autism spectrum disorder. The literature reveals a severely impaired functioning on all pragmatic aspects in autistic subjects, including difficulties in using speech acts (Ziatas, Durkin, & Pratt, 2003), comprehending irony and metaphor (Gold, Faust, & Goldstein, 2010; Martin & McDonald, 2004), detecting violations of Gricean Maxims (Surian, Baron-Cohen, & Van der Lely, 1996) and using features of context in utterance interpretation (Loukusa et al., 2007) (cited in Cummings, 2013, 2014). There is also clear evidence of mental state attribution impairments in autism (Baron-Cohen et al., 1985; Golan, Baron-Cohen, & Golan, 2008; Leslie & Frith, 1988; D. Williams & Happ, 2010; D. M. Williams & Happ, 2009). However, subjects with autism, despite communication deficits, do not seem to have difficulties in simulation, that is, suffering from limited imagination or difficulties in perspective taking.

Empirical research on whether autistic subjects are particularly susceptible to perspective taking impairments was first started by Hobson (1984). In one test, called 'seek-and-hide', subjects were presented a display which included hiding holes, a figure wishing to hide, and seekers. In a second test, the cube test, participants were requested to figure out the perspective of a doll, but in order to do so, they needed to recognize the relevance of points of view very different from their own. Results, as Hobson reports, show no deficiency in the ability to identify another subject's perspective, nor any difficulty to coordinate different perspectives in the cube game. The autistic participants performed as well as the comparison groups on both tasks. In a subsequent study, Leslie & Frith (1988) presented autistic children with a scene in which a plastic board was placed on a table such

that a doll could be on either side of the board (visible to the child). Next, a counter was introduced and placed on the board. Depending on the position of the counter in relation to the doll, the child was asked whether the doll could see the counter. In other trials, the experimenter varied the position of the doll while the child was asked to change the position of the counter where the doll could or could not see it. All of the participants passed the test, a result suggesting that subjects with autism have no difficulty to visualize the line of sight regardless of their ability to understand mental states.

Using a different paradigm, Baron-Cohen (1989) examined visual perspective taking in autistic subjects compared to the ability of normal subjects and subjects with Down's Syndrome. Small toys were placed around the subject and, from the orientation of an experimenter's eyes alone, the subject had to identify which toy the experimenter is attending to. The results show no significant differences within or between the three groups: 92.5% of autistic subjects, 94.4% of normal subjects, and 89.3% of subjects with Down Syndrome passed the test. This finding was replicated in a study by Leekam and colleagues (1997). Using the same procedure, Leekam et al. (Leekam, Baron-Cohen, Perrett, Milders, & Brown, 1997) compared the performance of autistic subjects relative to the performance of normal subjects and subjects with Down's Syndrome, and found no significant group differences.

The above finding all point in the direction of an intact capability of visual perspective taking, that individuals with autism can imaginatively understand that they and others might have a different line of sight, that what things others do and do not see at any given moment. This is what Flavell (1977) calls the ability of level-1 visual perspective taking. While the above results demonstrate that visualization in autism is intact, the mindreading exhibited in communication requires more than level-1 perspective taking. In addition to that, communication requires the ability of understanding that others may represent the same thing a bit differently than they do and thus might have a different perspective on the same thing. This is the ability Flavell (1977) called level-2 perspective taking.

Reed & Peterson (1990) examined both level-1 and level-2 perspective taking in autism. For level-2 perspective taking task, Reed and Peterson examined subjects' understanding of contrasting perspectives of individuals viewing the same object from different vantage points. The subjects sat in front of a turntable and an object (a plastic tiger or a teddy bear) was placed on the turntable. Participants were instructed to "turn it round so I can see the —" the last word being "nose," "tail," "back," depending on the object presented (Reed

& Peterson, 1990, p. 460). The authors report a uniformly high-level performance by all subjects on both level-1 and level-2 perspective taking task, with no significant difference in performance between autistic and control groups. Similarly, Tan & Harris (1991) found autistic subjects performing as well as normal controls on both level-1 and level-2 perspective taking tasks.

To examine visuospatial perspective taking in subjects with Asperger syndrome (autistic subjects with higher than average intellectual abilities), David et al. (2010) asked participants to detect an elevated object from a virtual character's point of view: "which object is elevated from his perspective?" Besides, participants went through an introductory session in which they were instructed to answer the questions by using imagination: "imagine yourself standing in the position of the virtual character"; "it is important to imagine your change in position!". David et al. report no significant group differences in perspective taking. The ability to spontaneously adopt visuospatial perspective in autism was also examined in a study by Zwickel et al. (Zwickel, White, Coniston, Senju, & Frith, 2011). A dot appeared next to a triangle protagonist, and participants were asked to press either the left or right button to indicate on which side of the screen the dot appeared relative to the triangle. On congruent trails, a dot appeared on the same side from the viewpoint of both the triangle and the observer, whereas on incongruent trials the dot occurred while the triangle was pointing downwards so a dot on the participant's right fell on the left of the triangle or vice versa. The results, as Zwickel et al. remark, were clear cut: subjects with autism and control participants displayed the same result, suggesting that spontaneous perspective taking in autism is intact.

Results regarding level-2 perspective taking have not always been consistent. Yirmiya et al. (Yirmiya, Sigman, & Zacks, 1994) presented autistic subjects with items on a rotating table. The task required the participants to turn the table until they can see an item in the exact same way that the experimenter can see the item from where she is standing. The majority of autistic subjects showed good perspective taking ability, but as a group, they performed less well than the normally developing children. However, the authors remark that the poor performance might be affected by the heavy memory demands of the test. So, poor performance in some cases may represent a difficulty in executive function than impairments in perspective taking. A study by Reed (2002) shows how, at least in some cases, deficits in executive function can explain poor performance in perspective taking.

Also, Hamilton et al. (Hamilton, Brindley, & Frith, 2009) compared the performance of autistic subjects on level-2 perspective tak-

ing to performance on a mental rotation task in a group of low-functioning children with autism. Both tasks began with the same experimental design: a toy placed on a turntable and the child was given a laminated card showing four pictures of the toy taken from different perspectives. For mental rotation trials, the toy was covered with an opaque flower pot, and the experimenter turned the table through 90°clockwise, and 180°and 90°counter clockwise. Then the child was asked, "when I lift the pot, which panda will you see?". For visual perspective taking task, the toy was covered and small doll, named Susan, was placed on different sides of the table and the child was asked: "this is Susan. When I lift the cover, which Panda will Susan see?". The results show that participants with autism have difficulties with level-2 visual perspective taking compared to their special rotation abilities. However, even the mental rotation task requires the subjects be able to imagine the rotation of the toy on the turntable.

Other sorts of evidence may help to resolve discrepant findings. One line of evidence comes from studies on first-person mindreading. If simulation is the mechanism that underlies mindreading—so subjects with autism have communication difficulties because they are impaired in perspective taking—then, as Carruthers (2006) has pointed out, on the assumption that introspective faculty is intact in autism, it should be expected that autistic subjects have no difficulty in reading their own minds. Several studies, however, speak against this assumption.

Phillips et al. (Phillips, Baron-Cohen, & Rutter, 1998) tested autistic subjects on an intention reporting task. The task was based on a target-shooting game in which the subjects' aim was to get as many prizes as possible by shooting down the cans with the prizes inside. Subjects had to choose a color to shoot, but before shooting a corresponding colored card was placed in front of the subject to remind the color he is shooting. Further, the test was designed in a way that the experimenters could manipulate the desirability of some outcomes, to make them intentional but not desirable, or desirable but accidental. As a result, some intended outcomes were not rewarded with a prize, and some accidental outcomes did contain a prize. After obtaining a prize, the subjects were asked "Which color did you mean to shoot? The (red) one or the (yellow) one?" The results show that when there was a discrepancy between intention and outcome (intended but not desirable or vice versa), subjects with autism performed much poorer than controls on reporting their intention, showing that subjects with autism hardly can discriminate between their intended and non-intended actions. In a more recent study, Williams & Happ (2010) assessed the explicit awareness children with autism

have of their own intentions. In experiment 1, participants awareness of their own knee-jerk reflex movements was assessed. Compared to controls, autistic subjects were significantly less accurate in reporting their reflex movements as unintentional. In experiment 2, the Transparent Intention task, participants with autism were less able than controls to recognize their mistaken action as unintended.

Using experience sampling method, Hurlburt et al. (Hurlburt, Happ, & Frith, 1994) studied introspection in three adult subjects with Asperger's syndrome. Subjects carry a small device that produces a beep at random intervals which the subject hears through an earphone. The subjects were instructed to freeze the content of their experience at the moment they hear the beep and then write down some notes about the details of that experience. Results by Hurlburt et al. show that one of the subjects, Peter, was unable to think and talk about his inner experience. The other two subjects also performed impressively different from normal controls, reported their thoughts purely in terms of images with almost no other features of experience such as emotional feelings.

Kazak et al. (Kazak, Collis, & Lewis, 1997) examined the understanding that subjects with autism have of their own knowledge state and that of another person and found no evidence showing that referring to ones own mental states in autistic subjects is easier than recognizing another persons mental states. Using a paradigm often labeled as unexpected content, Perner et al. (Perner, Frith, Leslie, & Leekam, 1989) presented autistic subjects with a typical box of a certain brand of sweets (Smarties), but to their surprise, the box contained something else (pencils). When asked about what they knew about the content of the box, the subjects had difficulty in reporting what they had wrongly predicted (Smarties) at the beginning of the experiment. In a more recent study, Williams et al. (Williams & Happ, 2009) devised an unexpected content task, the Plasters' task, in which the experimenter pretended that he had cut his finger and asked the participants if he could get him a plaster by pointing to the containers. The participants unexpectedly found that the plaster box contained birthday cake candles. Once discovered the actual content of the box, the participants were asked the Self-test question ("Before you looked in the tube, what did you think was inside?"), and the Other-person test question ("Later on I am going to show this tube to your teacher. He/she hasnt seen inside here though. What will he/she think is in there before he/she looks inside?"). The results show that autistic subjects have significant difficulty in reporting their prior false belief than to predict the false belief of another person.

The pattern we found in autism, that is, the dissociation between

communication deficits and perspective taking impairments, is also found in at least two other clinical populations with pragmatic disorders: patients with schizophrenia and fragile X syndrome.

Compared to subjects with mixed anxiety-depression and subjects with hemispheric brain damage, individuals with schizophrenia exhibit a high degree of inappropriate pragmatic abilities (Meilijson, Kasher, & Elizur, 2004). The impairments include proverb comprehension (Brne & Bodenstein, 2005), difficulties in processing contextual information (Bazin, Perruchet, Hardy-Bayle, & Feline, 2000; Sitnikova, Salisbury, Kuperberg, & Holcomb, 2002) and recognition of communicative intentions. Tnyi et al. (Tnyi, Herold, Szili, & Trixler, January February) presented patients with schizophrenia and normal control subjects with 'question and answer' vignettes, where the Gricean maxim of relevance was violated to communicate a hidden meaning. The findings show that schizophrenic patients fail to recognize the intentional violation of the Gricean maxim, a difficulty in understanding conversational implicatures. But Schizophrenic patients are also impaired in mental state attribution. Corcoran et al. (Corcoran, Cahill, & Frith, 1997) presented schizophrenic patients with two sets of cartoon jokes, the first set could be understood on the basis of purely physical and semantic analysis, while understanding the other set required participants to recognize the character's mental states in order to 'get' the joke. The patients had considerable difficulty to understand the jokes. Other studies also demonstrate difficulties in mental state attribution in schizophrenic patients (Bora, Yucel, & Pantelis, 2009; Brne & Bodenstein, 2005; Frith & Corcoran, 1996; Langdon et al., 1997).

However, subject with schizophrenia, similar to subjects with autism, do not seem to have difficulty in perspective taking. Langdon et al. (Langdon, Coltheart, Ward, & Catts, 2001) examined the visual perspective taking in schizophrenic subjects and normal controls. Participants were presented with arrays of colored blocks on a stand or a turntable and were asked two sets of questions. 'Item' questions (asking locations of array-features) and 'appearance' questions (asking how an array appear from another perspective), each type of question paired with viewer rotation instructions (asking subjects to imagine moving themselves relative to an array) and array-rotation instructions (asking subjects to imagine rotating an array relative to their fixed viewpoint). The results show that patients made more errors than controls only when judging appearance question, but performed as well as controls and with the same accuracy when judging all other questions.

Pragmatic impairments are also found in subjects with fragile X

syndrome. Specific pragmatic disturbances are observed in individuals with fragile X syndrome, including use of repetitive language (utterance repetition, topic repetition, and conversational device repetition) (Ferrier, Bashir, Meryash, Johnston, & Wolff, 1991; Murphy & Abbeduto, 2007), and failure to signal non-comprehension language as a listener (Abbeduto et al., 2008). Subjects with fragile X also show difficulty in mental state attribution (Cornish et al., 2005; Garner, Callias, & Turk, 1999; Grant, Apperly, & Oliver, 2007). However, fragile X subjects do not appear to have difficulty with perspective taking (Mazzocco, Pennington, & Hagerman, 1994; Mazzocco & Reiss, 1999).

So, to recap: simulation is not the mechanism responsible for the mindreading exhibited in communication for two reasons. First, not only simulation would be cognitively too demanding and impractical but also virtually ineffective in pragmatic interpretation. Second, as evidence from three clinical populations of autism, schizophrenia, and fragile X syndrome revealed, while there is an association between difficulties in mental state attribution and pragmatic deficits, perspective taking impairments and pragmatic deficits are dissociated. Thus, pragmatic comprehension cannot be simulational.

If pragmatic interpretation is not simulational, what is it then? This paper is not primarily intended to answer this question. However, despite the simulational proposals for utterance interpretation, most theorists working in pragmatics see utterance interpretation as a variety of mindreading and see mindreading itself guided and achieved either by a general-purpose inferential process (which operates during both utterance interpretation and other cases of mindreading) or by a module dedicated to the domain of utterance interpretation. Grice himself thought that utterance comprehension is an inferential process in which expression and recognition of communicative intentions are achieved by several levels of metarepresentation and application of a psychological theory (that is, the Gricean maxims and Cooperative Principle). Other theorists even have gone one step further arguing that the gap between what is said and what is communicated is so great that there is no way of establishing communicative intentions without a specialized module dedicated to utterance comprehension (Sperber & Wilson, 2002; Wilson, 2005). Now, while simulation might be the mindreading mechanism in other domains, for instance in recognition and attribution of emotions and sensations (Goldman, 2006; Goldman & Sripada, 2005) [14], given the cognitive complexity of mindreading exhibited in utterance comprehension and the empirical evidence we reviewed, simulation cannot be responsible for utterance

[14]For a critique of the role of simulation in emotion recognition, see (Yousefi Heris, 2017)

interpretation and prospects for a viable simulational account in this domain do not seem so promising.

0.5.4 Bibliography

Abbeduto, L., Murphy, M. M., Kover, S. T., Giles, N. D., Karadottir, S., Amman, A., ... Nollin, K. A. (2008). Signaling noncomprehension of language: a comparison of fragile X syndrome and Down syndrome. American Journal of Mental Retardation: AJMR, 113(3), 214-230.

Apperly, I. (2010). Mindreaders: The Cognitive Basis of Theory of Mind. Psychology Press.

Baron-Cohen, S. (1988). Social and pragmatic deficits in autism: Cognitive or affective? Journal of Autism and Developmental Disorders, 18(3), 379-402.

Baron-Cohen, S. (1989). Perceptual role taking and protodeclarative pointing in autism. British Journal of Developmental Psychology, 7(2), 113-127.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a theory of mind? Cognition, 21(1), 37–46.

Bazin, N., Perruchet, P., Hardy-Bayle, M. C., & Feline, A. (2000). Context-dependent information processing in patients with schizophrenia. Schizophrenia Research, 45(1–2), 93-101.

Bora, E., Yucel, M., & Pantelis, C. (2009). Theory of mind impairment in schizophrenia: Meta-analysis. Schizophrenia Research, 109(1-3), 1-9.

Botterill, G., & Carruthers, P. (1999). The Philosophy of Psychology. Cambridge, U.K.; New York: Cambridge University Press.

Brne, M., & Bodenstein, L. (2005). Proverb comprehension reconsidered-theory of mind and the pragmatic use of language in schizophrenia. Schizophrenia Research, 75(23), 233-239.

Carruthers, P. (2006). [Review of Review of Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading, by A. I. Goldman].

Corcoran, R., Cahill, C., & Frith, C. D. (1997). The appreciation of visual jokes in people with schizophrenia: a study of mentalizing ability. Schizophrenia Research, 24(3), 319-327.

Cornish, K., Burack, J. A., Rahman, A., Munir, F., Russo, N., & Grant, C. (2005). Theory of mind deficits in children with fragile X syndrome. Journal of Intellectual Disability Research: JIDR, 49(Pt 5), 372-378.

Cummings, L. (2013). Clinical Pragmatics and Theory of Mind. In

A. Capone, F. L. Piparo, & M. Carapezza (Eds.), Perspectives on Linguistic Pragmatics (pp. 23-56). Springer International Publishing.

Cummings, L. (2014). Pragmatic disorders and theory of mind. In L. Cummings (Ed.), The Cambridge handbook of communication disorders (pp. 559-577). Cambridge: Cambridge University Press.

Currie, G., & Ravenscroft, I. (2002). Recreative Minds: Imagination in Philosophy and Psychology. Oxford University Press.

David, N., Aumann, C., Bewernick, B. H., Santos, N. S., Lehnhardt, F.-G.,& Vogeley, K. (2010). Investigation of Mentalizing and Visuospatial Perspective Taking for Self and Other in Asperger Syndrome. Journal of Autism and Developmental Disorders, 40(3), 290-299.

Davies, M., & Stone, T. (1995a). Folk psychology: the theory of mind debate. Oxford; Cambridge, Mass.: Blackwell.

Davies, M., & Stone, T. (1995b). Mental Simulation: Evaluations and Applications - Reading in Mind and Language (1 edition). Oxford, UK; Cambridge, Mass: Wiley-Blackwell.

Ferrier, L. J., Bashir, A. S., Meryash, D. L., Johnston, J., & Wolff, P. (1991). Conversational Skills of Individuals with Fragile-X Syndrome: A Comparison with Autism and down Syndrome. Developmental Medicine & Child Neurology, 33(9), 776-788.

Flavell, J. H. (1977). The development of knowledge about visual perception. Nebraska Symposium on Motivation, 25, 43–76.

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., & Rizzolatti, G. (2005). Parietal Lobe: From Action Organization to Intention Understanding. Science, 308(5722), 662-667.

Frith, C. D., & Corcoran, R. (1996). Exploring theory of mind in people with schizophrenia. Psychological Medicine, 26(3), 521-530.

Gallese, V. (2001). The Shared Manifold Hypothesis: From Mirror Neurons to Empathy. Journal of Consciousness Studies, 8(5–7), 33-50.

Gallese, V., & Goldman, A. I. (1998). Mirror neurons and the simulation theory of mindreading. Trends in Cognitive Sciences, 2(12), 493-501.

Garner, C., Callias, M., & Turk, J. (1999). Executive function and theory of mind performance of boys with fragile-X syndrome. Jour-

nal of Intellectual Disability Research: JIDR, 43 (Pt 6), 466-474.

Golan, O., Baron-Cohen, S., & Golan, Y. (2008). The Reading the Mind in Films Task [Child Version]: Complex Emotion and Mental State Recognition in Children with and without Autism Spectrum Conditions. Journal of Autism and Developmental Disorders, 38(8), 1534–1541.

Gold, R., Faust, M., & Goldstein, A. (2010). Semantic integration during metaphor comprehension in Asperger syndrome. Brain and Language, 113(3), 124-134.

Goldman, A. I. (1989). Interpretation Psychologized. Mind & Language, 4(3), 161–185.

Goldman, A. I. (2006). Simulating Minds. Oxford University Press.

Goldman, A. I. (2008). Mirroring, Mindreading, and Simulation. In J. A. Pineda (Ed.), Mirror Neuron Systems (pp. 311-330). Humana Press.

Goldman, A. I. (2012). Theory of Mind. In E. Margolis & S. Stich (Eds.), The Oxford Handbook of Philosophy of Cognitive Science. Oxford University Press.

Goldman, A. I., & Jordan, L. (2013). Mindreading by Simulation: The Roles of Imagination and Mirroring. In S. Baron-Cohen, M. Lombardo, & H. Tager-Flusberg (Eds.), Understanding Other Minds: Perspectives from Developmental Social Neuroscience (pp. 448-466).

Goldman, A. I., & Sripada, C. S. (2005). Simulationist models of face-based emotion recognition. Cognition, 94(3), 193-213.

Gordon, R. M. (1986). Folk Psychology as Simulation. Mind & Language, 1(2), 158-171.

Grant, C. M., Apperly, I., & Oliver, C. (2007). Is theory of mind understanding impaired in males with fragile X syndrome? Journal of Abnormal Child Psychology, 35(1), 17-28. Grice, H. P. (1957). Meaning. The Philosophical Review, 66(3), 377.

Grice, H. P. (1969). Utterers Meaning and Intention. Philosophical Review, 78(2), 147–177. Reprinted in Grice, 1989.

Grice, H. P. (1989). Studies in the Way of Words. Harvard University Press.

Hamilton, A. F. de C., Brindley, R., & Frith, U. (2009). Visual perspective taking impairment in children with autistic spectrum disorder. Cognition, 113(1), 37-44.

Harris, P. L. (1992). From Simulation to Folk Psychology: The Case for Development. Mind & Language, 7(1-2), 120-144.

Heal, J. (1995). How to Think About Thinking. In M. Davies & T. Stone (Eds.), Mental Simulation. Blackwell.

Heal, J. (1996). Simulation and Cognitive Penetrability. Mind & Language, 11(1), 44-67.

Hobson, R. P. (1984). Early childhood autism and the question of egocentrism. Journal of Autism and Developmental Disorders, 14(1), 85-104.

Hurlburt, R. T., Happ, F., & Frith, U. (1994). Sampling the form of inner experience in three adults with Asperger syndrome. Psychological Medicine, 24(2), 385-395.

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., & Rizzolatti, G. (2005). Grasping the intentions of others with ones own mirror neuron system. PLoS Biology, 3(3), e79.

Kahneman, D., & Tversky, A. (1981). The Simulation Heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), Judgment Under Uncertainty: Heuristics and Biases. Cambridge University Press.

Kazak, S., Collis, G. M., & Lewis, V. (1997). Can Young People with Autism Refer to Knowledge States? Evidence from Their Understanding of Know and Guess. Journal of Child Psychology and Psychiatry, 38(8), 1001-1009.

Langdon, R., Coltheart, M., Ward, P. B., & Catts, S. V. (2001). Visual and cognitive perspective-taking impairments in schizophrenia: A failure of allocentric simulation? Cognitive Neuropsychiatry, 6(4), 241-269.

Langdon, R., Michie, P. T., Ward, P. B., McConaghy, N., Catts, S. V., & Coltheart, M. (1997). Defective Self and/or Other Mentalising in Schizophrenia: A Cognitive Neuropsychological Approach. Cognitive Neuropsychiatry, 2(3), 167-193.

Leekam, S., Baron-Cohen, S., Perrett, D., Milders, M., & Brown, S. (1997). Eye-direction detection: A dissociation between geometric and joint attention skills in autism. British Journal of Developmental Psychology, 15(1), 77-95.

Leslie, A. M. (1987). Pretense and representation: The origins of theory of mind. Psychological Review, 94(4), 412-426.

Leslie, A. M., & Frith, U. (1988). Autistic childrens understanding of seeing, knowing and believing. British Journal of Developmental

Psychology, 6(4), 315-324.

Loukusa, S., Leinonen, E., Kuusikko, S., Jussila, K., Mattila, M.-L., Ryder, N., ... Moilanen, I. (2007). Use of Context in Pragmatic Language Comprehension by Children with Asperger Syndrome or High-Functioning Autism. Journal of Autism and Developmental Disorders, 37(6), 1049-1059.

Martin, I., & McDonald, S. (2004). An exploration of causes of non-literal language problems in individuals with Asperger Syndrome. Journal of Autism and Developmental Disorders, 34(3), 311-328.

Mazzocco, M. M., Pennington, B. F., & Hagerman, R. J. (1994). Social cognition skills among females with fragile X. Journal of Autism and Developmental Disorders, 24(4), 473-485.

Mazzocco, M. M., & Reiss, A. L. (1999). A behavioral neurogenetics approach to understanding the fragile X syndrome: Contributions to a New Framework from the Cognitiv. In H. Tager- Flusberg(Ed.), Neurodevelopmental Disorders: Contributions to a New Framework from the Cognitiv (pp. 43-63). Cambridge, MA: MIT Press..

Meilijson, S. R., Kasher, A., & Elizur, A. (2004). Language Performance in Chronic Schizophrenia: A Pragmatic Approach. Journal of Speech, Language, and Hearing Research, 47(3), 695-713.

Murphy, M. M., & Abbeduto, L. (2007). Gender differences in repetitive language in fragile X syndrome. Journal of Intellectual Disability Research, 51(5), 387-400.

Nichols, S., & Stich, S. P. (2003). Mindreading. An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds. Oxford University Press.

Pearson, A., Ropar, D., & Hamilton, A. F. de C. (2013). A review of visual perspective taking in autism spectrum disorder. Frontiers in Human Neuroscience, 7.

Perner, J., Frith, U., Leslie, A. M., & Leekam, S. R. (1989). Exploration of the autistic childs theory of mind: knowledge, belief, and communication. Child Development, 60(3), 688-700.

Perner, J., & Khberger, A. (2005). Mental simulation: Royal road to other minds? In Other Minds (p. 166-181.). New York, Guilford Press.

Phillips, W., Baron-Cohen, S., & Rutter, M. (1998). Understanding intention in normal development and in autism. British Journal of

Developmental Psychology, 16(3), 337-348.

Reed, T. (2002). Visual Perspective Taking as a Measure of Working Memory in Participants With Autism. Journal of Developmental and Physical Disabilities, 14(1), 63-76.

Reed, T., & Peterson, C. (1990). A comparative study of autistic subjects performance at two levels of visual and cognitive perspective taking. Journal of Autism and Developmental Disorders, 20(4), 555-567.

Saxe, R., & Baron-Cohen, S. (Eds.). (2007). Theory of Mind: A Special Issue of Social Neuroscience (1 edition). Hove: Psychology Press.

Sitnikova, T., Salisbury, D. F., Kuperberg, G., & Holcomb, P. J. (2002). Electrophysiological insights into language processing in schizophrenia. Psychophysiology, 39(6), 851-860.

Sperber, D., & Wilson, D. (2002). Pragmatics, Modularity and Mind-reading. Mind & Language, 17(1-2), 3-23.

Stich, S. P. (2009). Stich and His Critics. (M. Bishop & D. Murphy, Eds.). Blackwell.

Stich, S. P., & Nichols, S. (1992). Folk Psychology: Simulation or Tacit Theory? Mind & Language, 7(1–2), 35-71.

Stich, S. P., & Nichols, S. (1995). Second thoughts on simulation. In Davies, Martin & T. Stone (Eds.), Mental Simulation: Evaluations and Applications.

Surian, L., Baron-Cohen, S., & Van der Lely, H. (1996). Are children with autism deaf to gricean maxims? Cognitive Neuropsychiatry, 1(1), 55-72.

Tan, J., & Harris, P. L. (1991). Autistic children understand seeing and wanting. Development and Psychopathology, 3(2), 163-174.

Tnyi, T., Herold, R., Szili, I. M., & Trixler, M. (January February). Schizophrenics Show a Failure in the Decoding of Violations of Conversational Implicatures. Psychopathology, 35(1), 25-27.

Wicker, B., Keysers, C., Plailly, J., Royet, J.-P., Gallese, V., & Rizzolatti, G. (2003). Both of Us Disgusted in My Insula: The Common Neural Basis of Seeing and Feeling Disgust. Neuron, 40(3), 655-664.

Williams, D., & Happ, F. (2010). Representing intentions in self and other: studies of autism and typical development. Developmental Science, 13(2), 307-319.

Williams, D. M., & Happ, F. (2009). What did I say? Versus what did I think? Attributing false beliefs to self amongst children with and without autism. Journal of Autism and Developmental Disorders, 39(6), 865-873.

Wilson, D. (2005). New directions for research on pragmatics and modularity. Lingua, 115(8), 1129-1146.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young childrens understanding of deception. Cognition, 13(1), 103-128.

Yirmiya, N., Sigman, M., & Zacks, D. (1994). Perceptual perspective-taking and seriation abilities in high-functioning children with autism. Development and Psychopathology, 6(2), 263–272.

Yousefi Heris, A. (2017). Why emotion recognition is not simulational. Philosophical Psychology, 0(0), 1-20.

Ziatas, K., Durkin, K., & Pratt, C. (2003). Differences in assertive speech acts produced by children with autism, Asperger syndrome, specific language impairment, and normal development. Development and Psychopathology, 15(1), 73-94.

Zwickel, J., White, S. J., Coniston, D., Senju, A., & Frith, U. (2011). Exploring the building blocks of social cognition: spontaneous agency perception and visual perspective taking in autism. Social Cognitive and Affective Neuroscience, 6(5), 564-571.

0.6 Willing, Intending, Metarepresenting: Weakness of Will Psychologized

Abstract

Philosophers have always treated weakness of will as acting against one's best judgment. But more recently Holton has suggested that weakness of will is agents failure to persist in resolutions. However, precisely what resolutions are and what forming resolutions cognitively involves are left rather underdescribed in Holton's account. In this paper, I first briefly clarify the concept of resolution and show that having a resolution involves the capacity of metarepresentation and intention recognition. This will provide a framework for a more natural and empirically oriented account of weakness of will. Next, I present substantial evidence from developmental psychology which, I argue, demonstrates that weakness or strength of will is essentially an exercise in metarepresentation.

0.6.1 Weakness of Will

An agent's action reveals weakness of will if he acts freely and intentionally counter to his own assessment of the action. In Davidson's words, in doing x an agent acts incontinently if and only if:

> (a) the agent does x intentionally; (b) the agent believes there is an alternative action y open to him; and (c) the agent judges that, all things considered, it would be better to do y than to do x. (Davidson, 1970, p. 22)

Condition (a) requires that the action is done consciously and deliberately. Condition (b) requires that the agent does the action from a range of options open to him at the time and so is doing that action freely. Condition (c) requires that the agent's overall assessment of the options open to him speak against doing x.

For centuries, this has been the dominant understanding of weakness of will in the philosophical literature. Philosophers have reacted to this picture in two different ways. Plato's Socrates, for instance, denied the possibility of weak-willed actions:

> no one who knows or believes there is something else better than what he is doing, something possible, will go on doing what he had been doing when he could be doing

> what is better. To give in to oneself is nothing other than ignorance, and to control oneself is nothing other than wisdom. (Protagoras, 358c)

Similarly, according to the moral philosopher R. M. Hare (1952), it is impossible for a person to act freely and intentionally against his moral judgment because value-judgments, if they are action-guiding, entail first-person commands or imperatives. On Hare's account, "it becomes analytic to say that everyone always does what he thinks he ought to" provided he is physically and psychologically able to do so (Hare, 1952, p. 169). Other philosophers, however, have tried to vindicate the possibility of weakness of will. For instance, Davidson (1970) holds that "there is no proving such actions exist; but it seems to me absolutely certain that they do" (Davidson, 1970, p. 29). It is possible to be incontinent and incontinent action is a real phenomenon; however, according to Davidson, the weak-willed man does not hold contradictory beliefs, but only violates the "principle of continence" and acts and judges irrationally.

As illustrated by this brief review, philosophers often have either denied or vindicated weakness of will's possibility. More recently, however, Holton (Holton, 1999, 2003, 2009) has developed a view by which he departs from almost all of the literature on the subject. On Holton's account, weakness of will is characterized "not as cases in which people act against their better judgment, but as cases in which they fail to act on their intentions" (Holton, 1999, p. 241). Suppose, Holton notes, there is some action I think I should perform in future, but I know when the time comes, I will be tempted not to do it. Then,

> it would be useful to form an intention now, an intention that will lead me directly to act when the time comes, and that will provide some resistance to reconsideration in the light of the inclinations I shall have then. Similarly, suppose that I know that my future reasoning will go awry: after a few glasses of wine my confidence in my own abilities will be absurdly high. Then again it would be good to form intentions now that are somewhat resistant to reconsideration in the light of those beliefs. In short, it would be good to have a specific type of intention that is designed to stand firm in the face of future contrary inclinations or beliefs: what I shall call a *resolution*. (Holton, 2009, p. 9–10)

An agent reveals weakness of will if he revises the intention he forms to overcome contrary beliefs and desires he expects to have. Weakness of will is not acting against one's best judgment, but failure to persist in

resolutions. Correspondingly, strength of will is ones success to persist with his resolutions (Holton, 2003, p.39).

I think a potential advantage of Holton's view is that it can move weakness of will from being considered as a problem about the puzzling nature of acting against one's best judgment into a more tractable problem about intention and resolution. This requires a clear conception of resolution. However, precisely what resolution is and what forming a resolution cognitively involves are left rather underdescribed in Holton's account. This is the task I take over in this paper. First, I briefly discuss the concept of resolution and show that having resolutions involves metarepresentational competence. I think this metarepresentational construal provides a framework which allows for a more natural and empirically oriented account of weakness of will. To that end, I review a body of evidence from developmental psychology which, I argue, supports the view that weakness or strength of will depends essentially on our metarepresentational skills.

0.6.2 Resolution: A New Psychological Construct

Central to Holton's account is the notion of irrational intention reconsideration. An agent reveals weakness of will if he readily revises an intention he has already formed to overcome his contrary desires and temptations. Following Bratman (1987) and Harman (1986), Holton makes a clear, non-reductive, distinction between beliefs, desires and intentions, and holds that intentions play a role that beliefs and desires could not. Intentions, unless they are revised, link agents to actions and are "controlling". Moreover, intentions have "stability" because once formed, they have a tendency to persist. Rethinking the previously formed intentions is too cognitively demanding, but intentions provide a way of storing decisions so we can act on them when the time comes without reconsideration (Holton, 1999, 2009). Whereas this stability makes intentions relatively immune to reconsideration, what is needed to stand firm against temptations cannot just be an intention, but a specific type of intention, what Holton calls resolution. Weak-willed people readily abandon their resolution and those with strong will-power stick to their resolutions even in the face of contrary desires. What precisely are resolutions, how they differ from intentions, and what is it that makes them firm against temptations? Holton notes:

> At one extreme we could think of them simply as intentions with a specially high degree of stability. But that doesn't seem to get it right. It is no part of the nature of a resolution that it will be effective; the point is rather that

> it is meant to be. At the most intellectual level, resolutions can be seen as involving both an intention to engage in a certain action, and a further intention not to let that intention be deflected. (Holton, 2009, p. 11)

Resolutions, which are thought of as intentions about a previously formed intention, are a species of second-order representation.[15]

We frequently, though often unconsciously, recognize and attribute mental states to others, but in addition, we are perfectly competent in detecting mental states in ourselves when, for instance, we become aware of our own desires, or when checking the consistency of our beliefs. This form of metarepresentation is often called metacognition, or first-person mindreading. Precisely what kinds of mechanisms underpin attribution (or detection) of mental states to self or others has been a matter of considerable debate over the last few decades, but much of this debate, for example, whether the underlying processes are theory-laden or simulation-style, or whether first- and third-person mindreading are subserved by parallel or dual processes, is orthogonal to our discussion here. Of the accounts that have been offered to explain the way we access our own mental states, the most promising one, in my view, is proposed by Nichols & Stich (2003) . On this account, metacognition involves the capacity (of a form) of introspection which operates when one is detecting and representing his own mental states. Nichols & Stich remark,

> the basic facts are that when normal adults believe that p, they can quickly and accurately form the belief *I believe that p*; when normal adults desire that p, they can quickly and accurately form the belief *I desire that p*; and so on for other basic propositional attitudes like intend and imagine. (Nichols & Stich, 2003, p. 160)

On this account, the capacity to form (first-person) metarepresentations is accomplished by a "Monitoring Mechanism" that takes, for instance, the belief representation p as input and returns metarepresentations of the form *I believe that p*.

To form resolutions, one must be able to represent his own intentions. This requires at least a two-step procedure: first, one must have an idea of his own intention. Having an idea here means that the person to be capable of classifying intentions as intention in a way that can recognize them as different from other related mental

[15]Belief, desires, and intentions are all first-order representations—representations directed towards the world, though with different conditions of satisfaction. Resolutions are second-order representations in which mind represent its own representations.

states such as beliefs and desires. To achieve this, there must be a mechanism operation of which brings an intention under the intention descriptor. What is required here is a Monitoring Mechanism (Nichols & Stich, 2003) that takes an intention as input, embed it into the relevant representational schema, and produce a representation R of his own intention that *I intend r*. This operation promises some self-awareness, and that the agent has an idea of his own intention. However, resolutions involve more than this. It is one thing to have an intention and another thing to have an intention about your intention. Forming resolutions involves a second step in which an intention is formed about an already recognized intention. This is carried out by a process in which Monitoring Mechanism takes R as input to generate representations of the form *I intend that R*, where R is the output of the previous step, and itself a metarepresentation.

Forming resolutions then relies on our ability of metarepresentation, indeed three representational stages of having an intention, recognizing (and becoming to some extent aware of) your intention, where this is distinct from recognizing other related mental states, and at last forming an intention about your intention. Understood this way, weakness of will turns into a problem in moral psychology, rather than moral philosophy, such that the plausibility of our understanding of weakness of will becomes very much dependent on what empirical literature reveals about our ability of metarepresentation, intention recognition compared to what is experimentally known about our ability of resisting temptations. But what do we empirically know about our ability to forgo immediate temptations? Experimental investigation on this ability began almost 50 years ago.

Mischel and colleagues developed the delay of gratification paradigm in which children were tested by creating a conflict between the temptation of taking an immediately available reward or waiting for a more preferred but delayed one. Children are typically seated in front of a table with a bell and rewards on it (e.g. marshmallows, two grouped and one apart). The child then faces a dilemma. The subject is told that the experimenter must leave the room for now, and she can get two marshmallows if she can wait for the experimenters return, but only one if she rings the bell and ends the delay. Results of an early study (Mischel & Mischel, 1987) show that 4-year-olds face serious difficulty to overcome temptations. However, within a year their performance in the test significantly improves mainly by rejecting arousing thoughts about temptations (Mischel & Mischel, 1987; Mischel, Shoda, & Rodriguez, 1989). Thompson et al. (Thompson, Barresi, & Moore, 1997) used a modified version of delay of gratification to test 3- to 5-year-olds in conflict situations, where the subjects had a choice

between two desirable alternatives: self-gratification now or forgoing current desires to gratify their own future desires, or the current or future desires of another child (shared gratification). Consistent with previous findings, results by Thompson et al. demonstrate that children younger than 4 show significantly less future-oriented prudence (benefits for self) and altruism (benefiting others) than older children.

Following previous studies, Kerr & Zelazo (2004) created a child variant of the Iowa Gambling task in which children chose between two decks of cards, where one deck offered more rewards (candies) per trial but were disadvantageous due to large losses, and the other deck contained fewer rewards but were advantageous overall. It is expected that over time and after several trials, when children learn about the decks, those who can resist temptations should be able to forgo immediate rewards and wait for better, but delayed options. Kerr & Zelazo found that children younger than 4 made more disadvantageous than would be expected by chance. Besides, they observed that younger children's performance did not improve across trials.

Using a similar version of the Iowa Gambling task, Garon & Moore (2007) examined one hundred and eighty-one children to assess age-related changes in performance on the gambling task, and association of the gambling task with a delay of gratification task. Garon & Moore found performance on the gambling task was significantly associated with performance on the delay of gratification task. The results confirmed previous findings by Kerr & Zelazo (2004); unlike 4.5-year-olds who showed a preference for the good deck, 3.5-year-old children selected more from the bad deck. Moreover, Garon & Moore tested children with awareness questions, asking them which deck was better or worse and why. Interestingly, they found that 3.5-year-olds, who could correctly tell which deck was best, failed to resist temptations and choose from this deck in the following block. In a second experiment, Garon & Moore (2007) explored the effect of labeling decks on children's performance. The labeling, while improved the performance of 4.5-year-olds, had no effect on the performance of younger children.

Overall, the evidence from these and other studies strongly suggests that the ability to resist temptations becomes more likely with increasing age and is almost impossible before 4 years of age. The crucial question now is how well this developmental outcome meshes with our understanding of weakness of will? If weakness of will is one's failure to persist in resolutions, then children at or above 4 years of age who, according to the above findings, resist temptations do so because they are in fact capable of forming resolutions, whereas younger children fail. Now, considering that resolutions involve metarepresentation and intention recognition, this requires:

(1) children who can resist temptations must be capable of metarepresentation and intention recognition, and
(2) younger children who fail to resist temptations also be hardly capable of metarepresentation and intention recognition.

However, it would be a serious challenge for the above understanding of weakness of will and our second-order construal of resolutions if it turns out that:

(3) children can resist temptations before the time they become capable of metarepresentation and intention recognition, or
(4) children fail to resist temptations even after the time they become capable of metarepresentation and intention recognition.

In what follows, I will argue that conditions (3) and (4) hardly square with what we find about childrens ability of metarepresentation and intention recognition. To support (1) and (2), I discuss a body of empirical evidence from developmental psychology which, I contend, demonstrates that younger children with difficulty to overcome temptations also fail to recognize intentions and represent mental representations, whereas children who show strength of will face no difficulty.

0.6.3 Weakness of Will Psychologized

It is one thing to have mental states and quite a different thing to understand and have a concept of mental states. To have a concept of belief requires understanding that beliefs are representations of reality, and not reality itself. Human mind represents objects and states of affairs in the world, but human adults also realize that (first-order) mental representations are themselves potential objects of representation. At what stages of normal human development we come to have a concept of mind—belief, desire, intention etc.—and understand the representational nature of mental states? The acid test for the presence of this ability is the false-belief task. In the original version of the task, Wimmer & Perner (1983) presented children a puppet show in which Maxi places his chocolate in the green cupboard and leaves the scene. A second puppet, Maxi's Mom, transfers the chocolate from the green cupboard to the blue cupboard. Children were then asked: when Maxi returns, where will he look for his chocolate? Wimmer and Perner found that some 4-year-olds and most 5-year-olds correctly predict that Maxi will look for the chocolate where she left it.

To give the correct answer, the child must be able to set aside his own representations of reality and think about what Maxi thinks about the chocolate's location. Children must realize that Maxi has a false belief, or misrepresentation, of reality. Besides, children who pass the test understand that people's beliefs determine how they act. That is why older children can predict Maxi's behavior correctly, whereas 3-year-olds and many 4-year-olds fail and wrongly believe Maxi will look for the marble where the marble actually is.

The results by Wimmer and Perner is supported by other research in this area. Flavell et al. (Flavell, Flavell, & Green, 1983) presented children a sponge painted to look like a rock. Children could then touch and squeeze the object to determine that the rock was, in fact, a sponge. Next children were asked two questions: "What is this really, really? Is it really, really a rock, or really, really a piece of sponge?" and "When you look at this with your eyes right now, does it look like a rock or does it look like a piece of sponge?" (Flavell et al., 1983, p. 102). Flavell et al. report that children younger than 4 responded that the object is a sponge, and it looks like a sponge, whereas older children answer in the way that adults do. This suggests that children younger than 4 do not understand mind's representational capacity: because they cannot make the appearance-reality distinction, they do not grasp how something that looks rock can, in fact, be a piece of sponge.

The results are corroborated by a variant form of the false belief test, labeled as unexpected content, in which children are shown a familiar container, often a tube of Smarties (popular British candy), and asked what is inside. Children often answer as expected: 'Smarties'. Next, it is revealed that the real content is something quite different, pencils for example. The child faces the test question: what someone else, the child who is outside the room and has not seen inside the box, will think is in the box. To answer correctly, children must set aside their own belief of the box and predict the answer by thinking what the other child thinks. Studies (Hogrefe, Wimmer, & Perner, 1986; Perner, Leekam, & Wimmer, 1987) show that children younger than 4 attribute their own representation of the real content to the other child, even though in response to the Ignorance Question [16] they correctly diagnosed that the other child did not know the real content of the box.

If, as traditionally thought by some philosophers, we have a privi-

[16]The Ignorance Question asks "Does [name of the other child] know what is really in the box, or does he [she] not know that?". The Belief Question asks "If we ask [name of the other child], what will he [she] say is in the box?" (Hogrefe, Wimmer, & Perner, 1986, p. 569).

leged access to our own mind, then maybe representing our own mental representations must be easier and developmentally earlier than representing other minds. This view, however, came under attack by philosophers (Sellars, 1956) and empirical research on the theory of mind. To examine the ability to understand representational change in oneself, Gopnik and Astington (1988) asked children: "When you first saw the box, before we opened it, what did you think was inside it? Did you think there were Smarties inside it or did you thing there were pencils inside it?" (Gopnik & Astington, 1988, p. 29). In correspondence with previous findings, the results show that children under 4 barely appreciate representational change and that one's past and present representations may contrast.

A considerable concern about the above findings is that the false-belief test might underestimate children's metarepresentational competence. One difficulty with the test is that the child is never told the protagonist's belief, but, from what the child sees, he or she must infer that Maxi still believes the chocolate is in its original location. This probably makes the test unnecessarily difficult. It would be cognitively less demanding if the child is told directly about the protagonist's belief. Wellman & Bartsch (1988) created a simplified version of the test in which children were told: "Jane's kitten is really in the playroom, but Jane thinks the kitten is in the kitchen. Where will Jane look for her kitten?" (Wellman & Bartsch, 1988, p. 264). The results show that 3- and many 4-year-olds fail to predict the action in terms of where the object was believed to be. This suggests children's difficulty in the original test were not belief-inference related but stems from a failure in representing target's relevant belief even in an extremely simplified version of the test.

Could the task be linguistically demanding or misleading? Do children understand the test question in the way adults do? If not, younger children's failure demonstrates, in fact, a failure to understand what is communicated than a metarepresentational difficulty. The test question that "where will Maxi look for his chocolate?" may simply be interpreted as "where will Maxi have to look for his chocolate?" rather than the intended question that "where will Maxi will look for his chocolate *first*?". Clements & Perner (1994) modified the test question to explicitly communicate which point in time is being asked in the question. The results show no significant difference in childrens responses whether the word 'first'was used (36% passing) or not (32% passing). Similarly, Gopnik & Astington (1988) obtained no difference in children's performance when included a wider range of syntactic forms of the question, and Moses & Flavell (1990) found younger children would not do better even in very rich context when

the task is made easy by giving children very strong belief cues.

Different studies have tried easier methodological variants of the test, but results from a meta-analysis of 178 false-belief studies by Wellman and colleagues (Wellman, Cross, & Watson, 2001), which examined effects of those variables that might influence childrens performance, revealed that no set of manipulations can enhance younger children's performance to above chance. The meta-analysis study and other evidence all seem to point to the conclusion that younger childrens performance in the false-belief test demonstrates a conceptual, metarepresentational difficulty beyond any difficulties related to task requirements.

This conclusion, however, has been called into question by a growing body of more recent research on infants and younger children. Using indirect behavioral methods, rather than explicit judgments, several studies report a dissociation between performance on a verbal measure and behavioral measure on the false-belief test. Clements and Perner (1994) found that most 3-year-olds, despite making incorrect judgments on the verbal measure, show a looking pattern consistent with the correct (initial) location of the object. Several other studies (Garnham & Perner, 2001; Garnham & Ruffman, 2001; Low, 2010) report that, despite providing incorrect verbal prediction, children looked in anticipation to the correct location. Besides, using methods other than eye tracking, several studies have obtained similar results for even younger children. O'Neill (1996) examined whether 2-year-old children take into account their parents mental state during communication. The child was introduced to a toy that was placed in a container on a high shelf. When asking for help in retrieving the toy, children significantly more often named the toy, named its location, and gestured to its location when a parent had not witnessed these events than when she or he had, suggesting that children even at the age of 2 modify their behavior according to what they think their parents think.

Using the violation-of-expectation paradigm, several studies claim for the metarepresentational competence even in younger children. Southgate, Senju, & Csibra (2007) present data which, according to the authors, strongly suggest that 25-month-olds make anticipatory saccades on the basis of false belief attribution. Onishi and Baillargeon (2005) found that 15-month-old babies look significantly longer when the protagonist's action was inconsistent with his false belief than when it was consistent. Buttelmann, Carpenter, & Tomasello (2009) obtained similar results for 18-month-old infants, and results from Surian, Caldi, and Sperber (2007) pushes the age down to 13 months. The findings are confirmed by results from other studies as well (Scott

& Baillargeon, 2009; Song, Onishi, Baillargeon, & Fisher, 2008; Truble, Marinovi, & Pauen, 2010).

Does the evidence show metarepresentational competence in stages that are developmentally earlier than the time obtained from previous studies? I think not. Several comments about the interpretation of these findings are in order. First, almost all of the evidence for early metarepresentational ability comes from looking-time paradigm, but what exactly can be inferred from looking time? The question has been controversial (Haith, 1998; Kagan, 2008). The essential premise of the looking paradigm is that infants look longer at unexpected events than those that are expected, but why should we suppose that prolonged looking-time demonstrates high-level cognitive processing, for instance, that the infants expected the protagonist to search for the toy according to his false belief? Prolonged looking time does demonstrate a behavioral disruption, or that some pattern is detected, but it is less than clear why this finding should count as an indicator of high-level cognitive processing such as belief, reasoning or metarepresentation. Besides, even failure to show a more looking time to an event does not mean that the infant experienced no expectancy violation. In a study by Kagan et al. (Kagan, Linn, Mount, Reznick, & Hiatt, 1979), infants did not show increased attention to a novel stimulus even when the old and new stimuli were known to be discriminable. In addition, results from looking-time pattern has not always been consistent: whereas Onishi and Baillargeon (2005) report violation of expectation in 15-month-old infants, Clements and Perner (1994) found no such effect in children younger than about 3.

Second, earlier metarepresentational competence is not definitely implied by the evidence because there is a rival, comparatively deflationary, interpretation of the findings that would not warrant this conclusion. The rival interpretation would say younger children's performance in these experiments can be based on behavioral rules (Perner & Ruffman, 2005). In the experiment by Onishi and Baillargeon (2005), what is observed is that 15-month-old infants expect the protagonist to act in a particular way. This expectation, however, can be based on a behavioral rule, in this case, infants may innately be disposed to assume that agents search for an object where they last saw it and not necessarily in its actual place. Based on behavioral rules, infants form a set of expectations and make predictions (from behavior to behavior) requiring no conception of the mind. The findings by Onishi and Baillargeon is compatible with the assumption that 15-month-old infants take into account the protaganist's mental state, but it is also equally compatible with the assumption that infants predict future behavior from current behavior without understanding mental states

as the mediating step. It is more parsimonious to assume that younger children take into account only observable behaviors. One might object that this criticism applies to older children too, but adults and children who pass the false-belief test demonstrate metarepresentional competence in a variety of situations, and only the presence of this ability can explain why they make correct predictions at different situations (Penn & Povinelli, 2007; Perner & Ruffman, 2005; Povinelli & Giambrone, 1999; Povinelli & Vonk, 2004).

To be clear, the claim is not that younger children show no understaning of mental states, but they have general difficulty in metarepresentation. Besides, even if younger children understand mental representations, they are not aware of this understanding. When children give an incorrect verbal answer, but look correctly to where the protagonist would search for the object, do they have conscious knowledge of where the character would search, or they are unaware of this possibility? To examine the question, Ruffman and colleagues (Ruffman, Garnham, Import, & Connolly, 2001) asked children to bet the counters on where they think the protagonist would search for the object. If eye gaze indexes conscious knowledge, children should bet counters on where they look. However, Ruffman et al. found that children who showed correct eye gaze, but an incorrect explicit verbal answer, bet very highly on the location consistent with their explicit answer. This finding not only suggests that young children are very confident about their (wrong) explicit answer, but seem completely unaware of the knowledge they convey through their eye gaze.

Resolutions, however, involve more than an unconscious understanding of mental states. If an agent decides to overcome a desire, for example, to resist the temptation of smoking a cigarette, the agent not only must be able to represent this intention, but also to be aware, at least to some degree, of having this intention, especially if the agent is going to form a second intention—that is, to form a resolution—about this previously formed intention.[17] To achieve this, the agent not only must be able to represent mental representations but also can recognize particular representations as intention. Can children recognize intentions before the time they can resist temptations?

I argued that children who fail to resist temptations also fail to represent mental representations. I now argue that weak-willed children fail to recognize intentions either. Several studies show recognition of intentions and goal-directed actions in very young children. In a study by Woodward (1998), infants were habituated to an event in

[17]mental state representation, even in adults, often occurs below the threshold of consciousness, but the knowledge is such that the agents are often can bring it to the conscious level upon request.

which an actor reached for and grasped one of two toys side by side on a stage. Following habituation, the toy positions were reversed and infants observed test events in which there was a change either in the toy the actor reached or the path of reach taken by the actor. Infants looked longer on new goal/old path trails than on old goal/new path trials, suggesting that infants at 6 months of age recognized actors intention and expected the agent to pursue the same goal. Using the habituation paradigm, other studies have tried to show infants' ability to detect goal-directed actions at 3 months of age (Sommerville, Woodward, & Needham, 2005), attribute goals to non-human agents by 3- and 5-month olds (Luo, 2011; Luo & Baillargeon, 2005), recognize the goal-directedness of actions at 10 and 12 months of age (Brandone & Wellman, 2009), and take the intentional stance by the first year of life (Csibra, Br, Kos, & Gergely, 2003; Gergely & Csibra, 2003; Gergely, Ndasdy, Csibra, & Br, 1995).

In a study by Meltzof (1995), 18-month-olds were confronted with an adult who merely demonstrated an intention to act in a certain way. The adult tried, but failed, to perform the action (for example, the adult tried to pull apart a toy, but failed) and children never saw the end state. It was hypothesized that children who recognize intention read through the body movements to the underlying intention of the action. Otherwise, they would interpret behavior in purely physical terms. Meltzof found that, when given a chance to act on the objects themselves, 18-month-olds reproduced the acts the adults intended to do even though the adult failed. The result, according to Meltzoff, suggests that children differentiate between the surface, physical behavior and a deeper level of intention recognition. Carpenter et al. (Carpenter, Akhtar, & Tomasello, 1998) found a similar pattern in children who were, on average, two months younger than subjects in the study by Meltzoff. Do the above findings suggest children recognize intentions before the time they can form resolutions or resist temptations? I think not.

To begin with, the findings on infants' understanding of intention relies mainly on habituation/dishabituation paradigm. Infants repeatedly observe a stimulus that embodies some conceptual principle, but once habituated, infants are confronted with test events in which the conceptual principle is violated. If infants exhibit a greater visual attention, many researchers conclude that the subjects possess an understanding of the concept instantiated by original stimulus. This, however, hardly seems to be the best interpretation of the evidence. According to some theorists (Baird & Baldwin, 2001; Povinelli, 2001), the dishabituation effect may reflect the operation of low-level mechanisms that detect physical and temporal regularities in

actions, and enable observers to identify relevant units in the behavior stream. As Povinelli remarks, "the early detection of structural regularities of behavior are not, strictly speaking, the early manifestation of the uniquely human system for reasoning about intentions" (Povinelli, 2001, pp. 240-241). Under this interpretation, understanding intentions requires a higher-level, psychological, mechanism which enables observers to make sense of the units in the behavior stream, but the higher- and low-level systems are evolutionarily and developmentally dissociable. So infants' detection of behavioral regularities does not necessarily imply any understanding of intention.

Besides, because the above studies rely on the assumption that intentional actions are cognitively harder to understand than non-intentional ones, many researchers feel comfortable to attribute intention understanding to children once they observe subjects' ability to recognize intentional actions. However, studies have shown that children and adults quickly develop a default (implicit) explanatory bias, where subjects' analysis and understanding of all actions are judged to be intentional by default (Kelemen, 1999; Kelemen & Rosset, 2009; Rosset, 2008). If so, contrary to the above assumption, judging actions as intentional is easier and requires less cognitive resources than judging them as non-intentional. Thus, apart from investigating when children understand intentions, research especially on young children must also take into account children's understanding of non-intentional actions. This view has been recently developed by Rosset & Rottman (2014), on which children's judgments are by default intentionally driven, but this default over-attribution of intention is gradually tapered off and inhibited through children's experience with non-intentional actions.

Now, even if the above evidence implies some understanding of intention—admitting that intention is not an all-or-nothing concept, but is acquired gradually—the presence of this understanding is limited to the ways children perceive actions and observable world. It is limited and close to what Searle (1983) calls intentions-in-action, a sense of intention which requires no prior plan, reasoning or any decision, but arises spontaneously when one is engaged in bodily movement. However, forming resolution involves shifting attention from actions and bodily movements to what Searle (1983) calls prior intentions. What is required is a form of intention agents have prior to performing an action and understand it as the mental cause of action.

When do children understand intentions as the mental cause of actions? To answer this question, Shultz and colleagues (Shultz, Wells, & Sarda, 1980) examined children's ability to distinguish between intentional and mistaken actions. Children were asked to perform a

number of tasks each of which contained an intentional behavior and an analogous behavior designed to constitute a mistake, for example, in repeating tongue twisters, or picking up objects (e.g. shiny penny) when subjects' vision was distorted with a set of prism glasses. After the trials, children were asked: "Did you mean to do that?". The results show that children as young as 3 judged that they did not mean to do what they did during the mistaken trials. However, a considerable concern about this and findings using similar strategy is that, to answer the questions correctly, children can only use a matching rule in which information about the stated goal or desire is matched against the behavioral outcome. If the desire and the outcome match, then it can be concluded that the outcome was intended, and if there is a mismatch, then the outcome can be judged as not intended (Astington & Lee, 1991; Shultz & Wells, 1985). When children pick up the shiny penny, the outcome satisfies their desire or stated goal and children judge the action as intended, but during the mistaken trails, when there is a mismatch between the desire and the outcome, the action is judged as not intended.

To understand intentions as the mental cause of actions, children must recognize intentions as different from other related mental states such as desires. To examine this, we need to look at situations where the matching strategy could not be used, for instance, in situations where an actor's desire or goal is never stated, or in experimental designs in which desire is fulfilled but intention is not. This condition is met in an early study by Smith (1978) in which subjects, while remained unaware of the actors desire or goal, watched a series of videotapes in which a young woman performed voluntary (arm exercises, sitting down on a chair etc.) and involuntary (yawning, sneezing etc.) actions. Smith found that children at 4 and younger judged all voluntary and involuntary actions as intentional, suggesting their failure to understand the causal link between intentions and actions. In a different study, Astington & Lee (1991) presented children with a pair of stories, in one of them a girl takes some bread outside, throws some crumbs down, and birds peck them up. In a second story, another girl takes some bread outside, but it just happens that some crumbs drop behind the girl, and birds peck them up. Children were asked, "Which girl meant the birds to eat the crumbs?". Astington and Lee found that only 5-year-olds performed quite well, with 3-year-olds performing only at chance level.

Similarly, to prevent the use of matching strategy, Feinfield et al. (Feinfield, Lee, Flavell, Green, & Flavell, 1999) presented three- and four-year-olds with illustrated stories in which the story characters' intention differed both from their desires and from the outcomes: char-

acters wanted to go to location A (friend's house) but decided to go to a place B they disliked (skating rink) because their mother wanted them to go there. However, because the bus driver gets lost, they unintentionally ended up at location A where they really wanted to go. Children were then given three test questions: (1) Where did X try to go–location B or A? (2) Remember when X was deciding where to go? Where did he think he was gonna go– location A or B? (3) Where does X like to go–location B or A? To show the ability of identifying characters intention, children had to distinguish intention from the agent's desire and from the actual outcome, both of which different from the agent's intention. Feinfield et al. (1999) found that, unlike the four-year-olds, the three-year-olds performed worse than would be expected by chance.

Because the story-comprehension tasks like those in Feinfield et al.s are difficult for younger children—requires the effort of encoding the narrative of the story—Schult (2002) created a target-hitting game which consisted of tossing beanbags into colored buckets. The purpose of the study was to examine children's ability to identify and distinguish their own intentions and desires. In this game, children could either hit the intended target or a different target, and could win a prize or not win a prize in a way that the desire to win the prize could be satisfied independently of the intention to hit a particular target. Children then faced to sets of "identification" (What you were trying to do?) and "satisfaction" questions (Did you do what you were trying to do?). If the children understood intentions, they should have been able to report which target they were trying to hit regardless of the outcome of their throw (desired or not). For the identification questions, the results show that, unlike 4- and 5-year-olds, the 3-year-olds could not report their intention. When 3-year-olds missed their intended target, but satisfied by finding a prize, they said they were aiming for that unintended target all along, suggesting that children at this age confuse intentions and desires and cannot differentiate them in a mismatch condition.

Let us recap what we have done so far. Having clarified the concept of resolution, the central claim of the paper was that weakness or strength of will crucially depends on our metarepresentational skills. To support this, I presented substantial evidence to show that there is a solid link between weakness of will and the abilities of metarepresentation and intention recognition. I argued that evidence from developmental psychology suggests agents who experience difficulty to resist temptations also face difficulty to recognize intentions and represent mental representations and agents who show strength of will do so only after the time they become capable of metarepresentation and inten-

tion recognition. If this is correct, we should also expect agents with an underdeveloped ability of metarepresentation be hardly capable of resisting temptations. We should expect that, for instance, autistic children, who are known as having metarepresentational deficits, face serious difficulty to resist temptations. Interestingly, this prediction turns out to be correct. Faja & Dawson (2013) compared 21 children with autistic spectrum disorder (ASD) and 21 typically developing children between 6 and 7 years of age on a delay of gratification task. Results from direct observation and parent report show that ASD children, who were intellectually at or above the average range, were incapable of waiting to receive a more desired reward.

As seen, a clear advantage of the view presented here, compared to the traditional understanding of acting against one's best judgment, is that the view not only accounts for a wide range of empirical findings but provides a framework which allows for a more natural and psychologically oriented understanding of the problem. Rather than to rely merely on conceptual analysis, the plausibility of the view can be assessed against further empirical evidence.

Now, is weakness of will possible? Unlike the traditional view, we need not discuss, deny or vindicate weakness of will's possibility. It is part of our life and experience. The crucial question rather is *how* it happens and that what mechanisms underlying its occurrence. I have not offered a complete solution to this question; neither have I thought Holton's account provide an adequate answer to this question. In my view, weakness of will is probably the outcome of several interacting factors but, given the evidence we reviewed above, metarepresentation and intention recognition are among the most crucial, perhaps the most crucial, factors that must be taken into account in future research.

0.6.4 Bibliography

Astington, J. W., & Lee, E. (1991). What do children know about intentional causation. SRCD, Seattle.

Baird, J. A., & Baldwin, D. A. (2001). Making sense of human behavior: Action parsing and intentional inference. Intentions and Intentionality: Foundations of Social Cognition, 193-206.

Brandone, A. C., & Wellman, H. M. (2009). You cant always get what you want: Infants understand failed goal-directed actions. Psychological Science, 20(1), 85-91.

Bratman, M. (1987). Intention, Plans, and Practical Reason. Cambridge: Harvard University Press.

Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. Cognition, 112(2), 337-342.

Carpenter, M., Akhtar, N., & Tomasello, M. (1998). Fourteen- through 18-month-old infants differentially imitate intentional and accidental actions. Infant Behavior and Development, 21(2), 315-330.

Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. Cognitive Development, 9(4), 377-395.

Csibra, G., Br, S., Kos, O., & Gergely, G. (2003). One-year-old infants use teleological representations of actions productively. Cognitive Science, 27(1), 111-133.

Davidson, D. (1970). How Is Weakness of the Will Possible? In Essays on Actions and Events (Oxford University Press, 1980) (pp. 21-42).

Faja, S., & Dawson, G. (2013). Reduced delay of gratification and effortful control among young children with autism spectrum disorders. Autism: The International Journal of Research and Practice.

Feinfield, K. A., Lee, P. P., Flavell, E. R., Green, F. L., & Flavell, J. H. (1999). Young Childrens Understanding of Intention. Cognitive Development, 14(3), 463-486.

Flavell, J. H., Flavell, E. R., & Green, F. L. (1983). Development of the appearance-reality distinction. Cognitive Psychology, 15(1), 95-120.

Garnham, W. A., & Perner, J. (2001). Actions really do speak louder than wordsbut only implicitly: Young childrens understanding of

false belief in action. British Journal of Developmental Psychology, 19(3), 413-432.

Garnham, W. A., & Ruffman, T. (2001). Doesnt see, doesnt know: is anticipatory looking really related to understanding or belief? Developmental Science, 4(1), 94 -100.

Garon, N., & Moore, C. (2007). Awareness and Symbol Use Improves Future-Oriented Decision Making in Preschoolers. Developmental Neuropsychology, 31(1), 39-59.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: the nave theory of rational action. Trends in Cognitive Sciences, 7(7), 287-292.

Gergely, G., Ndasdy, Z., Csibra, G., & Br, S. (1995). Taking the intentional stance at 12 months of age. Cognition, 56(2), 165-193.

Gopnik, A., & Astington, J. W. (1988). Children's Understanding of Representational Change and Its Relation to the Understanding of False Belief and the Appearance-Reality Distinction. Child Development, 59(1), 26.

Haith, M. M. (1998). Who put the cog in infant cognition? Is rich interpretation too costly? Infant Behavior and Development, 21(2), 167-179.

Hare, R. M. (1952). The Language of Morals. Oxford Clarendon Press.

Harman, G. (1986). Change in View. MIT Press.

Hogrefe, G.-J., Wimmer, H., & Perner, J. (1986). Ignorance versus False Belief: A Developmental Lag in Attribution of Epistemic States. Child Development, 57(3), 567-582.

Holton, R. (1999). Intention and Weakness of Will. The Journal of Philosophy, 96(5), 241-262.

Holton, R. (2003). How is Strength of Will Possible? In C. Tappolet & S. Stroud (Eds.), Weakness of Will and Practical Irrationality (pp. 39-67). Oxford.

Holton, R. (2009). Willing, Wanting, Waiting. Oxford: Oxford University Press.

Kagan, J. (2008). In Defense of Qualitative Changes in Development. Child Development, 79(6), 1606-1624.

Kagan, J., Linn, S., Mount, R., Reznick, J. S., & Hiatt, S. (1979). Asymmetry of inference in the dishabituation paradigm. Canadian

Journal of Psychology, 33(4), 288-305.

Kelemen, D. (1999). Why are rocks pointy? Children's preference for teleological explanations of the natural world. Developmental Psychology, 35(6), 1440-1452.

Kelemen, D., & Rosset, E. (2009). The Human Function Compunction: Teleological explanation in adults. Cognition, 111(1), 138-143.

Kerr, A., & Zelazo, P. D. (2004). Development of hot executive function: The childrens gambling task. Brain and Cognition, 55(1), 148-157.

Low, J. (2010). Preschoolers' Implicit and Explicit False-Belief Understanding: Relations With Complex Syntactical Mastery. Child Development, 81(2), 597-615.

Luo, Y. (2011). Three-month-old infants attribute goals to a non–human agent. Developmental Science, 14(2), 453-460.

Luo, Y., & Baillargeon, R. (2005). Can a Self-Propelled Box Have a Goal? Psychological Science, 16(8), 601-608.

Mischel, H. N., & Mischel, W. (1987). The Development of Children's Knowledge of Self-Control Strategies. In F. Halisch & J. Kuhl (Eds.), Motivation, Intention, and Volition (pp. 321–336). Springer Berlin Heidelberg.

Mischel, W., Shoda, Y., & Rodriguez, M. I. (1989). Delay of gratification in children. Science, 244(4907), 933-938.

Moses, L. J., & Flavell, J. H. (1990). Inferring False Beliefs from Actions and Reactions. Child Development, 61(4), 929-945.

Nichols, S., & Stich, S. P. (2003). Mindreading. An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds. Oxford University Press.

O'Neill, D. K. (1996). Two-year-old children's sensitivity to a parent's knowledge state when making requests. Child development, 659-677.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-Month-Old Infants Understand False Beliefs? Science, 308(5719), 255-258.

Penn, D. C., & Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a theory of mind. Philosophical Transactions of the Royal Society B: Biological Sciences, 362(1480), 731-744.

Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. British Journal of Developmental Psychology, 5(2), 125-137.

Perner, J., & Ruffman, T. (2005). Infants Insight into the Mind: How Deep? Science, 308(5719), 214-216.

Plato. Protagoras, in Plato: Complete Works, J. M. Cooper & D. S. Hutchinson, (Eds.), Indianapolis, Ind: Hackett Publishing Co., 1997, pp. 746-790.

Povinelli, D. J. (2001). On the possibilities of detecting intentions prior to understanding them. Intentions and Intentionality: Foundations of Social Cognition, 225-248.

Povinelli, D. J., & Giambrone, S. (1999). Inferring Other Minds: Failure of the Argument by Analogy. Philosophical Topics, 27(1), 167-201.

Povinelli, D. J., & Vonk, J. (2004). We Don't Need a Microscope to Explore the Chimpanzees Mind. Mind & Language, 19(1), 1-28.

Rosset, E. (2008). It's no accident: Our bias for intentional explanations. Cognition, 108(3), 771-780.

Rosset, E., & Rottman, J. (2014). The Big "Whoops! "it is raining" in the Study of Intentional Behavior: An Appeal for a New Framework in Understanding Human Actions. Journal of Cognition and Culture, 14(1-2), 27-39.

Ruffman, T., Garnham, W., Import, A., & Connolly, D. (2001). Does Eye Gaze Indicate Implicit Knowledge of False Belief? Charting Transitions in Knowledge. Journal of Experimental Child Psychology, 80(3), 201-224.

Schult, C. A. (2002). Children's Understanding of the Distinction between Intentions and Desires. Child Development, 73(6), 1727-1747.

Scott, R. M., & Baillargeon, R. (2009). Which Penguin Is This? Attributing False Beliefs About Object Identity at 18Months. Child Development, 80(4), 1172-1196.

Sellars, W. S. (1956). Empiricism and the Philosophy of Mind. Minnesota Studies in the Philosophy of Science, 1, 253-329.

Shultz, T. R., & Wells, D. (1985). Judging the intentionality of action-outcomes. Developmental Psychology, 21(1), 83-89.

Shultz, T. R., Wells, D., & Sarda, M. (1980). Development of the ability to distinguish intended actions from mistakes, reflexes, and

passive movements. British Journal of Social and Clinical Psychology, 19(4), 301-310.

Smith, M. C. (1978). Cognizing the Behavior Stream: The Recognition of Intentional Action. Child Development, 49(3), 736-743.

Sommerville, J. A., Woodward, A. L., & Needham, A. (2005). Action experience alters 3-month-old infants perception of others' actions. Cognition, 96(1), B1-11.

Song, H., Onishi, K. H., Baillargeon, R., & Fisher, C. (2008). Can an agent's false belief be corrected by an appropriate communication? Psychological reasoning in 18-month-old infants. Cognition, 109(3), 295-315.

Southgate, V., Senju, A., & Csibra, G. (2007). Action Anticipation Through Attribution of False Belief by 2-Year-Olds. Psychological Science, 18(7), 587-592.

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of Beliefs by 13-Month-Old Infants. Psychological Science, 18(7), 580-586.

Thompson, C., Barresi, J., & Moore, C. (1997). The development of future-oriented prudence and altruism in preschoolers. Cognitive Development, 12(2), 199-212.

Truble, B., Marinovi, V., & Pauen, S. (2010). Early Theory of Mind Competencies: Do Infants Understand Others Beliefs? Infancy, 15(4), 434-444.

Wellman, H. M., & Bartsch, K. (1988). Young children's reasoning about beliefs. Cognition, 30(3), 239-277.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-Analysis of Theory-of-Mind Development: The Truth about False Belief. Child Development, 72(3), 655-684.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young childrens understanding of deception. Cognition, 13(1), 103-128.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actors reach. Cognition, 69(1), 1-34.

0.7 Conclusion

Simulation can succeed as long as there is an isomorphism between the simulating system and target. Otherwise, how am I justified in attributing mental states by simulation if I come to know that my mind or brain treats mental states in a way that is significantly different from the targets mind? But if the assumption of isomorphism is so crucial for simulation, what are the relevant respects of similarity in simulation? I have argued that all the potential candidates–phenomenological, functional, and neurological–run into intractable problems. Phenomenological and functional similarity are not sufficient for simulation, and neurological similarity fails to discriminate simulation from non-simulational processes. The problem can be solved if ST theorists could provide a theory of function, but I have argued that mirror neurons do not operate according to the standard simulation prototype. This faces the simulation hypothesis with a thorny dilemma: either we keep off-line simulation within the bounds of high-level mindreading and define simulation in terms of neural resemblance only, in which case we face serious problems, or, to avoid the problems, we generalize off-line simulation to low-level mindreading, in which case the operation of mirror neurons would not qualify as simulational.

This conclusion might seem too narrow and constraining, especially if we are convinced by arguments for neurological similarity. However, I have shown that the most celebrated cases of simulation at low-level mindreading fail to satisfy the two key proposed similarity requirements.

But if simulation processes do not show neurological similarity, on what grounds might the claim that mirror neurons are evidence in support of ST be justified? Mirror neurons support ST only under the narrow and localized interpretation of neuronal processing; however, I have shown that a closer examination of evidence demonstrates that a cognitive function, in particular, emotion recognition, occurs in a network of brain structures involving multiple cross-regional interactions in the brain. This suggests that, contrary to what ST theorists argue, mirror neurons are not incompatible with TT, the information-rich understanding of mindreading.

Mindreading plays significant roles in a variety of cognitive domains. In pragmatics, most theorists and researchers working today agree that people in communicative contexts are deeply involved in metal state attribution, either for what they trying to express or in recognizing what is expressed. If mindreading is simulational, then it is expected that the ST account can explain the mindreading exhibited

in communication. Contrary to this assumption, I have argued that simulation in this domain would be cognitively too demanding and virtually ineffective in pragmatic interpretation. In addition, I have shown that, while evidence demonstrates an association between difficulties in mental state attribution and pragmatic deficits, results from studies on three clinical populations of autism, schizophrenia, and fragile X syndrome reveals a dissociation between pragmatic deficits and simulation. This demonstrates that the underlying mechanisms of high-level mindreading, at least in the domain of utterance interpretation, cannot be simulational.

Further, I have suggested a link between mindreading and weakness, or strength, of will. How is it that people fail or succeed in resisting temptations? I propose that weakness or strength of will strongly depends on our ability of first-person mindreading, or metarepresentation, indeed three representational stages of having an intention, recognizing intention, and forming an intention about your intention. This suggestion turns weakness of will into a problem in moral psychology, rather than moral philosophy, in which we need not discuss, deny or vindicate weakness of wills possibility, as philosophers have traditionally done. Under this account, weakness of will is not only possible but it is an undeniable part of our life and experience. Rather, the crucial question now is how it happens and what psychological/neural mechanisms underpin its occurrence.

Questions concerning mindreading has been debated for more than two decades, but unlike the early debates which were more like a two-sided battle between theory and simulation, more recent theorists consider mindreading as a complex phenomenon which takes elements of both theory and simulation. This approach has been fruitful, adopted by many others and has resulted in several hybrid theories. Yet a different strategy in following this approach would be to evaluate the explanatory power of theory or simulation with respect to the roles mindreading plays in different domains and in connection with various cognitive functions. I have tried this approach in pragmatics, moral psychology, and emotion recognition. It remains for future research to explore and assess mindreading in other domains.

0.8 Curriculum Vitae

Education

- 2013-Present — University of Munich, Germany
  Graduate School of Systemic Neurosciences
- 2010-2012 — University of Leuven (KU Leuven), Belgium
  M.Sc. in Cognitive Science
- 2005-2007 — University of Liverpool, England, UK
  M.A. in Philosophy

Areas of Specialization

- Philosophy of Psychology, Philosophy of Mind, Philosophy of Neuroscience, Moral Psychology

Areas of Competence

- Moral Philosophy, Philosophy of Science, Epistemology

Presentations

- Weakness of Will: Resisting Temptations: Neurophilosophy Workshop San Servolo, Venice, Italy (2014)

0.9 Publications

- (2017) Yousefi Heris, A. "Why Emotion Recognition Is Not Simulational" *Philosophical Psychology*
- (2011) "On Concepts: With or Without Perceptions", *Philosophy Pathways, issue 167*

Publications Under Review

- Simulation, Mirroring, and Neurological Similarity
- Mindreading: The Concept of Simulation
- Mindreading: The Concept of Simulation
- Willing, Intending, Metarepresenting: Weakness of Will Psychologized

0.10 Eidesstattliche Versicherung/Affidavit

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation "Reading through Mirror Neurons? Mindreading Reconsidered" selbststndig angefertigt habe, mich auer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annhernd ubernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

I hereby confirm that the dissertation "Reading through Mirror Neurons? Mindreading Reconsidered"is the result of my own work and that I have only used sources or materials listed and specified in the dissertation.

Munich, May 2017

Ali Yousefi Heris

0.11 Declaration of Author Contributions

I declare that this thesis has been composed solely by myself, that the work contained herein is the result of my own work, and that this work has not been submitted for any other degree or professional qualification.

Part of this work, section 0.4, has been published in *Philosophical Psychology*.

22 May 2017

Ali Yousefi Heris Prof. Dr. Stephan Sellmaier (Supervisor)