12-2016

An Investigation into Hybrid Models of Mindreading: A Dual Type Theory Account

Alexandra Jewell
University of Arkansas, Fayetteville

An Investigation into Hybrid Models of Mindreading: A Dual Type Theory Account

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Arts in Philosophy

by

Alexandra Duree Jewell
University of Arkansas
Bachelor of Arts in Psychology, 2014

December 2016
University of Arkansas

This thesis is approved for recommendation to the Graduate Council.

_____
Dr. Jack Lyons
Thesis Director

_____
Eric M. Funkhouser
Committee Member

_____
Warren Herold
Committee Member

Abstract

Mindreading, or attributing mental states to others, involves instances of simulation and theory; but there is controversy over which one of these methods is the primary, or default, mechanism. I propose that mindreading is a theory-based process, such that we utilize theory over simulation when both are available and reliable. To argue my position, I suggest that theory has been inaccurately portrayed in past discussion and that we possess two types: a connectionist network (tt1) and a traditional, conceptual folk-psychology (tt2). By dividing theory in this way, we can explain common phenomena of mindreading that other theory-based accounts do not explicitly imply. Previously used as evidence for a simulation-based model, these phenomena are now compatible with a Dual Type Theory Account. Additionally, the distinction between type 1 and type 2 theory invites a new argument for primacy by appealing to the cognitive resources required by a mechanism, such that the primary method will be the one that requires the least amount of effort and is available in every case. Since tt1 is effortless and automatic, it is likely the default process. Tt1 provides us with a modular, fast, unconscious mindreading tool that is not dependent on conceptual knowledge, yet it can be influenced and adjusted by tt2 via supervised learning.

Keywords: mindreading, theory, simulation, hybrid account, default mechanism,

Table of Contents

An Investigation into Hybrid Models of Mindreading:
A Dual Type Theory Account

Human beings have the capability to attribute emotions and beliefs to others with decent accuracy, such that we can interact successfully in our environment most of the time. For example, we can correctly identify when an individual is upset and even assign reasons behind another's action[i]; and this attribution influences our beliefs and behaviors toward the other person. But what mechanism is underlying this phenomenon and how might that mechanism be playing its role? In fact, there are most likely different tools we can utilize to accomplish this feat, different ways of concluding some state which we believe another to have. Typically, there are two types of mechanisms: those which incorporate theory and those that do not. The former group fall under Theory Theory (henceforth TT), whereas the latter are generally members of Simulation Theory (henceforth ST). While classically it was believed that humans use one type of mechanism, with philosophers advocating for it exclusively over the other, currently there is growing popularity in holding a hybrid model, in which an individual utilizes both types of mechanisms in different situations; but even in these hybrid accounts, one mechanism is favored, or considered default, over the other— meaning it is primary, chosen most often, chosen when all mechanisms are available. I also hold a hybrid model and aim to show, by introducing a Dual Type Theory[ii] Account, that our default mechanism for mindreading is theory. I believe this new account can explain aspects of the phenomenon, which previously have been cited as evidence for a simulation-based account, as well as approach this discussion from a new perspective to determine the default mechanism.

While there are many hybrid models, mine differs from others by distinguishing between two types of theory, that is two types of mechanism used to mindread that utilize a theory. I

believe this distinction might illuminate theory as the default mechanism that humans use to attribute high-level mental states, like complex emotions[iii], desires, and beliefs, etc. to others. The inaccurate grouping of both types of theory into one might be a contributing factor that leads some authors to favor simulation as default. Some aspects cannot be explained by appealing to one type of theory alone, but by identifying the two types we can explain evidence originally understood as supporting simulation. These phenomena are no longer the strong evidence that simulation-advocates argue it is since it is also compatible with a Dual Type Theory Account. Additionally, I will provide reasons why we should believe theory to be the primary tool for this task: some reasons that, to my best knowledge, have not been addressed before[iv].

I will begin by discussing what is meant by mindreading, theory, and simulation according to the literature as well as mentioning relevant topics that will support my view. By identifying the related debates will we see the potential problems that have led me to hold that two types of theory exist and that one is default to the other and simulation. However, both theories along with instances of simulation can work together and have the ability to increase the accuracy of our mindreading capacity. We will examine some current models of hybrid accounts and unpack what authors mean by suggesting one is default, or primary, above another mechanism. We shall see that some of the reasons that authors like Goldman use to argue simulation as primary are not enough to support this claim since they are compatible on my view and could even be used to argue the opposite.

I will discuss the ways in which humans make these judgments regarding the mental states of others, the underlying mechanisms, and the connections that exist among said mechanisms. It is my goal to demonstrate that a hybrid model based on the Dual Type Theory Account explains different features of the mindreading phenomenon, such as the developmental

aspect, egocentric bias, and even instances of implicit bias; and it is also compatible with what we know about humans being cognitive misers and the demands on cognitive resources of different processes (to be discussed later). Within this account I will argue that type 1 theory (henceforth tt1) is, in fact, our default mechanism when assigning mental states to others because it is the mechanism that requires the least amount of resources, followed by the other tools: simulation and type 2 theory (henceforth tt2). Because I am dissecting theory in this way, I ought to discuss the connection between these different processes and the influence tt2 has on tt1, in addition to the role simulation plays within this phenomenon. When distinguishing between different types of processes and examining the characteristics that accompany the distinction, I will suggest that simulation either is a type 2 process or does not fit precisely into these categories: regardless, this suggestion might imply that we need to clarify our definitions or that these processes' features are not as clustered as previously believed.

Terminology

Mindreading

Mindreading, or 'Theory of Mind,' is the cognitive ability to attribute mental states to others, that is, the way in which people conclude what another person is feeling or believing. This phenomenon is pervasive in human activity as we must assume what others are thinking or feeling to interact with them appropriately. How might we know to comfort someone's sadness unless we have some way of attributing that mental state based on her appearance or situation via a mechanism that incorporates such information in order to conclude what she is likely feeling? Also, we care to explain or predict others' behavior since it might affect us.

We come to know our own mental states through introspection; we are capable of identifying our current emotions and beliefs because this type of knowledge is transparent to us[v].

Often, reflection can illuminate one's reasons for behaving a certain way or feeling a certain way. But the mental states of another are not readily available like those of our own. Nevertheless, we still engage in trying to identify and assign mental states to others because we know that other people are like us, and just as we have mental states, they must have beliefs, desires, and emotions that can explain behavior.

There are many instances of mindreading, that is, there are many cases of attributing various mental states to others, such that humans must have different mechanisms that can complete the varying tasks. Because these cases are so different we should take a hybrid account, one that incorporates simulation and theory, seriously. For example, to attribute an emotion like sadness to someone is very different than attributing a belief about what a person will do or what she believes, yet both are cases of mindreading. If an individual is seen with disappointed eyes and lips curled down into a frown, most of us would believe her to be sad and attribute sadness based on her appearance. But this case is different than attributing some mental state to her such as a belief, desire, or consequential action. In cases of beliefs, etc., appearance alone does not suffice in giving us the relevant information to accurately attribute the target's mental state. I am interested in the mechanism used in the latter type of mindreading, one that incorporates more than just appearances and is capable of assigning higher-level representational states to others. Because types of cases are so different, most theory of mind proposals have two or more systems or levels; however, the way to separate between the types differs among models. Distinguishing among types of mindreading is crucial in identifying which mechanisms are responsible for attributing different kinds of mental states; a mechanism used for one level might not be default for the other level. We can start by analyzing the division offered by Goldman (2006), which distinguishes between low-level and high-level simulation based on nonpropositional vs.

propositional content. Although this distinction is used to explain the differences specifically in terms of simulation, we can use the qualifications of his bi-level model to understand mindreading in general, despite the mechanism we believe to be responsible.

Similar to Goldman's partition is that of Tager-Flusberg & Sullivan (2000), which distinguishes between a "social cognitive" and a "social perceptual" component. The cognitive includes understanding the mind as a representational system, whereas the latter involves *perception* of biological and intentional behavior and recognition of emotions from facial expressions. These components map onto Goldman's high-level and low-level systems respectively (Goldman & Jordan, 2013). Low-level mindreading, which is stimulus driven, consists of mindreading nonpropositional states, such as sadness. The other level, high-level mindreading, is concerned with attitudes, beliefs, desires, and other propositional states, and this kind is the focus of this discussion. What process is default when attributing mental states such as beliefs to others? My view is more in line with the separation of Tager-Flusberg & Sullivan, rather than Goldman, since I am arguing that high-level is still stimulus-driven in a way, in that my default mechanism automatically occurs in the presence of a stimulus; but unlike low-level processes, it is concerned with beliefs, desires, etc. The mechanism incorporates more than mimicking perceptual cues of a target (as Goldman describes low-level simulation). High-level type of mindreading seems very different than the low-level or surface level attributions (propositional vs. nonpropositional), therefore this difference may lead be a good reason to believe the mechanisms behind the different kinds of attributions are not the same.

As aforementioned, the phenomenon of mindreading is complex; and there is growing support for hybrid theories, which include some cases being carried out by TT type, information-rich processes while some are simulated by information-poor processes[vi]. I also hold a hybrid

account, and so believe there are instances where simulation is the mechanism used to mindread, however I will argue that theory is default for high-level attribution. Even though it is not the focus of this paper, I would like to point out that low-level mindreading is one such instance in which simulation (something other than theory) is the tool for attribution. Goldman (2006) discusses the mirror-based mechanism responsible for low-level mindreading, such that when perceiving a person who is sad and exhibiting certain facial expressions, mirror neurons aid in replicating the exact state to accurately attribute nonpropositional states of this kind.

The overwhelming evidence of mirror-neurons and the studies that demonstrate this type of low-level mindreading suggest that Goldman is correct in proposing that such mindreading- nonpropositional, stimulus driven, perceptual- is a result of simulation via mirroring. When we see someone exhibiting sad facial expressions, we mimic those expressions, which ultimately leads us to conclude the sadness we attribute to the target. We unconsciously recreate the exact state we attribute to her by simulating the same state, which we see. Within my hybrid account, I do not reject that low-level mindreading is executed via mirror neurons, which is a type of simulation. However, I disagree with Goldman that our use of mirror neurons to mindread low-level cases is strong support for simulation of different kinds (specifically imagination) being default for high-level mindreading (2006). The cases are so different. It is not obvious that a non-theory mechanism or mimicking mechanism must be the tool used in both. While I concede that low-level mindreading is obtained via simulation (mirror neurons), I disagree that simulation (imagination) is the default mechanism behind high-level mindreading. Both instances of simulation aim at recreating the state of the target; but the differences in low-level and high-level cannot support a non-theory mechanism being default for high-level just because it is the tool in low-level mindreading. To better understand the default process of high-level mindreading, we

shall focus on the specifics of each mechanism and identity which one is the primary, automatic tool to attribute propositional states to others, such that we successfully interact with each other and the environment.

Theory

The dominant view for explaining most cognitive abilities is to propose an internal knowledge or theory, a set of rules that guides cognition; this similar strategy was believed to be responsible for mindreading (Fodor 1987, Sellars, 1963; as cited in Stich & Nichols, 1992). In this view, the internally represented knowledge, known as "folk-psychology," which is partly accessible to consciousness but more often not, is the underlying mechanism responsible for guiding our attributions of mental states and predictions of behavior to others. This knowledge is not concerning our own mental processes but those of other people. Some who hold this view consider the information used in mindreading to be similar to the knowledge structure found in a scientific theory and is acquired and utilized in much of the same way (Stich & Nichols, 2003). When I refer to "folk-psychology" for the remainder of this discussion, I am referring to the traditional, sentence-like, rule-based, internally represented, conceptual knowledge of other people's psychology. This traditional type of account is information-rich since it relies on vast knowledge of the minds of others, that is, in order for an individual to attribute certain mental states to others, she has to possess general knowledge of human psychology and have concepts pertaining to the mental states and beliefs that others might have. However, this dependence on an information-rich theory is a downfall of traditional theory, and I propose an alternative account of theory, such that this constant reliance on conceptual knowledge is not necessary.

A potential problem that arises when considering traditional TT to be the only framework available to individuals is childhood error especially the false-belief task[vii], such that a child

inaccurately attributes belief states to others based on the attributor's knowledge, which the target does not possess, but that she does not disregard. This type of problem also addresses the question of acquisition, how a child comes to possess this internally represented information, which changes as a child develops. That is that the concepts children have regarding the mental states of others gain sophistication as the child gets older, and some of these childhood errors decrease as she gets older. Two solutions from TT proponents include the child scientist theory and the modularity theory. According to the former, children are like scientist, who refine their theories during childhood and acquire folk-psychology like other learned scientific theories. In response to childhood error, as children gain more theories and concepts they begin to replace the non-representational theory with a representational one that allows them to conceptualize and avoid the false-belief problem. Modularity theorists propose that children are not developing or refining their theory as child-scientist theorists state, but rather the theory is innate of 'modules' and limitedly interacts with information from other components (Baron-Cohen, 1995; Scholl & Leslie, 1999). Since the theory is innate, these theorists explain the error as a result of the selection-processor failing to inhibit truths about the world when selecting inputs for the inference mechanism, but as a child develops she gets better at inhibition. I propose a modular-type theory but non-nativist; a theory that exhibits other aspects of modules, like automaticity and effortlessness, but it can modularize new information. Humans have an innate capacity to mindread; but modulation can also occur where learning, both from environmental inputs and conceptual knowledge, can overlearn our modules, and new connections become automatic. This type of learning will be discussed in further detail.

How might the mechanism based on traditional TT proceed? As previously stated, the more seasoned models include knowledge structures, on which the attributor relies when

ascribing mental states to others. The reason for the popularity of this type of model stems from similar ones explaining other cognitive capacities through explicit rules or sentence-like principles (Stich & Nichols, 1992). Goldman and Jordan (2013) provide a description of what mindreading via traditional TT might include when attributing a decision or plan to some target. For example, to predict which donut shop Donna will go to, a person incorporates beliefs attributed to Donna as well as aspects of general human psychology (folk-psychology) and feeds those inputs into the attributor's own theoretical reasoning mechanism or inference mechanism. If we know that Donna desires donuts, and she believes the best donuts are from DonutDisturb, we then possess two beliefs specifically about Donna. More precisely, we possess two metarepresentational beliefs about Donna, that is, we have beliefs about what Donna believes. Additionally, we have a general belief about human psychology that tells us that people will most likely do the thing that will satisfy their desires. When we feed these three beliefs into our inference or theoretical reasoning mechanism, we are left with the output that Donna will go to DonutDisturb, and thus designate this future behavior to her. See Figure 1. An important aspect of this TT process is that the attributable states are metarepresentational belief states of the target. The metarepresentational characteristic of this state is important when we consider self-reflection, its role in tagging a state as metarepresentational, and its occurrence during mindreading tasks (to be discussed later). Additionally important, the processor relied upon to produce the conclusion or the attributable mental state is only the theoretical reasoning mechanism or inference mechanism of the attributor.

The inference mechanism used to extract the relevant information from our internalized theory of mind is similar to that of other theories like folk-physics. We have such mechanisms that are capable of computing the internalized theories like folk-physics, so it is plausible that the

same tool can be used with respect to our folk-psychology and selects the elements to be considered during mindreading, that is, the relevant theories. However, while traditional TT has a mechanism "for free," a mechanism that we have available and use often, it relies on some encoded or represented generalizations or regularities of folk-psychology (Stich & Nichols, 1992). This reliance on elaborate information is a common critique for traditional TT, as simulationists often attack it as being too complex and offer simulation as a simpler explanation; however, simulation may not be as easy as they claim[viii].

If mindreading does occur through accessing internalized information regarding other peoples' minds, then the accuracy of the attribution relies on the folk-psychology on which the attributor appeals to. If the attributor utilizes a bad theory of human psychology, then she will be more likely to produce an incorrect conclusion and thus an inaccurate reading of the target. Say for example that in her folk-psychology an individual has the belief that a person will choose the closest option instead of the desired option to satisfy her needs. If this belief is fed into the theoretical reasoning mechanism along with the beliefs specific to Donna, then the attributor may conclude that Donna will go to DonutHut, which is significantly lower in appeal and taste but closer to Donna. If this later belief about general human psychology is false, then the attribution has a higher potential to be inaccurate. Therefore, the content of the knowledge relied upon within the TT framework guides the accuracy of the attribution.

Connectionist Networks. There has been a divergence from the sentence-like idea of knowledge structure, and instead a connectionist model has increased support when considering the internal database (Churchland, 1981 & 1989). Oversimplifying a great deal, connectionism is the view that the mind has networks consisting of many units, or nodes, and the connections among them depend on the weight or strength between the nodes[ix], so that certain ones are

activated given the activation others. Connectionist models state that the weights among the nodes play a role when making predictions based on this information, that is, the knowledge of the regularities of the world are represented by the strengths between nodes, such that we do not need a sentence-like theory to represent the world. I will not be offering arguments for this type of processing but such arguments have been given, for example, by Smolensky (1988b) and Chalmers (1990).

These two different portrayals of the knowledge structures may cause confusion when considering what should be included as "theory." For example, if theory is meant to suggest sentence-like rules or principles, then the connectionist network would not be included in a true internally represented theory of knowledge. On the other hand, if we use "theory" in a wide or liberal sense, then any type of internal information concerning a certain subject should encode a tacit theory (Stich & Nichols, 1992). I prefer to consider "theory" in this wide sense, and thus propose a theory of internal knowledge that includes connections of nodes and weights among them that influence which bits of knowledge are more salient or relied upon more often.

As opposed to conceptual representation like that in the traditional theory or folk-psychology, connectionists claim that information is stored within the weights, or connection strengths, between the nodes of a network; and the information stored as weights is non-symbolic. These types of neural nets do not follow strict rules like that of propositional, traditional theory, instead they respond to statistical patterns. It is not obvious that this type of information, the weighted nodes that connect the input and output, is accessible to consciousness, but it is certainly nonconceptual in general. If our mindreading is composed of a connectionist network, the presence of a folk-psychology becomes questionable. Again, on the wider interpretation of "theory" we are not limited to the sentence-like representation; and other

theory-theorists have cited the connectionist networks as included in what they mean by stating "theory." However, if the connectionist networks include aspects that are nonconceptual then perhaps we do not possess a conceptual folk-psychology, instead the theory is just a modular response based on nonconceptual weighted nodes that connect the input (say for instance an observed behavior) and some output (attributed desire). A folk-psychology is not necessary considering the network reflects the environment by adjusting weights, which are all nonconceptual. For example, if we witness a person hit another, we will most likely attribute anger to the aggressor. How do we go about making that attribution according to the different ideas of theory? If we use sentence-like, propositional beliefs concerning the psychology of others, we probably refer to our belief that "most people who do violence are angry." However, if we use some connectionist network, which relies on the weights that are based on probability given past experience, then maybe we do not need that belief at all. The belief is unnecessary because our connectionist network automatically reaches the likely output of anger considering the inputs, and the network has been wired to reliably reflect instances of violence and anger. What I mean is that our network is wired such that it is not necessary to access some symbolically-represented, propositional knowledge because the automatic, connectionist network already reflects that knowledge. When we saw instances of anger in the past, there was violence. And when we saw violence, we found that it was caused by anger. And so, the probability of anger being the cause of the perceived violence increased within our network; thus, in future cases of violence the output to be attributed will be anger. The positive weight between the nodes increased, so that the knowledge that anger causes violence rests in the strength of the nodes between the input (violence) and the output (anger); and the theory that anger causes violence is represented by that connectionist network.

If the knowledge or theory is encoded, in response to regularities in the environment, within the weights or strengths among the nodes in the connectionist network, then it is nonconceptual, specifically state-nonconceptual. It still represents the world, only the subject does not have to have the concepts needed to express the content (Evans, 1982). Evans (1982) argues that nonconceptual mental states are unconscious but become conscious when they serve as inputs to another system, such as a reasoning system; this is consistent with the idea that type 1 processes, like tt1, operate unconsciously, and we are only aware of the output of the processing (Bargh, 1994). It seems that this is the case with intuitive responses regarding mindreading; the intermediate nodes of the network are not consciously accessible to us, but the final outputs are because they are attributed to the target[x].

The nonconceptuality of the connectionist nodes needs to be addressed. Based on the distinction between state and content nonconceptuality from Heck (2000), the latter claims that the content of propositional attitudes and beliefs must be a different type of content than that of nonpropositional states; on the other hand, the former does not require more than one type of content. A connectionist network responsive to statistical probabilities and not dependent on the concepts a person possesses, could still potentially share content with that of propositional states, and will, in such a case, be state nonconceptual. I am going to assume that since the system can undergo the process without necessarily possessing the concepts, it is state nonconceptual; but I think that the content could be specified if the person possessed the concept, so I am not going to conclude content nonconceptuality from state nonconceptuality as some other authors do (Heck, 2007).

A network that is developed in a bottom-up, or stimulus-driven, way is independent of concepts and would be both synchronically and diachronically state nonconceptual, at no point is

it relying on concepts to be in the state that it is in. The fact that it is stimulus-driven does not entail that it applies to low-level mindreading because the information it represents reflects people's beliefs, desires, etc. Although typically nonconceptual, if the connectionist networks of our mindreading module (tt1) are susceptible to supervised learning, by which conceptual information (tt2) alters the weights among the nodes in a given network by making relevant information available to the module, then tt1 would be dependent on concepts; and therefore, the connectionist network theory is synchronically state-nonconceptual but diachronically conceptual.

There could be propositional information which is the same as, or could accurately specify, the information stored in the weights among nodes, hence why it is state and not content nonconceptual. For example, in the violence and anger connection as stated above, one could hold the belief that anger causes people to be violent while also having a connection network that generates the output "anger" when it has the input "violence." And since the content of the network and the propositional knowledge can overlap, if the network is susceptible to supervised learning, then it is possible that conceptual information may come into play within the connectionist network. While connectionist networks do not have to rely on concepts, the theory of connectionist networks that I am proposing can be affected by conceptual information, which might alter the strengths among the nodes at some point.

Some might argue (Eliminativists) that connectionism is responsible for mental states as opposed to folk-psychology, and beliefs and desires don't really exist, thus that folk-psychology doesn't exist (Churchland, 1981 & 1989). Again, a "folk-psychology," as I am using it, is the base of conceptual knowledge concerning the mental states and psychological processes of other people. This is not knowledge regarding my own mind, but the minds of others, and is usually

defined as internally represented information. However, I do not think this is the case; I do not think we have to sacrifice one theory for another. By identifying the nodes of connectionist networks as nonconceptual, as nets that respond to statistical patterns, they lose the necessity of concepts, that is, they lose the conceptuality aspect of the symbolic representation that is needed to possess a folk-psychology: conceptual information regarding the minds of others, but they retain the information or knowledge. Even though this folk-psychology is not present in the neural net, I think we still possess it.

The knowledge stored nonconceptually in the neural nets of tt1, can also be symbolically represented as conceptual, propositional knowledge regarding the minds of others in tt2. One reason why I believe we have this conceptual information is our ability to use it to change our connectionist networks. As stated before, connectionist nets are synchronically state-nonconceptual, but conceptual information can guide which nodes are strengthened by selecting the correct output for a certain input and making that knowledge available to tt1[xi]. In this way, networks depend on the concept while the weights are being tuned over many trials. The altered network reflects the chosen propositional information by the strengths between the input and output (again, think of the violence and anger example); and since this type of development of the connectionist network occurred via conceptual information, the network is now diachronically state conceptual. According to Stich and Nichols (1992), predictions of behavior and other high-level attributions are 'cognitively penetrable,' meaning they can be influenced by new theories learned by the attributor. They offer an example in which people are asked what sample they would prefer if given two of the same item. Some people, unaware of a certain psychological trend[xii], answered that they would choose at random; however, this is not what happens during an actual experiment. The wrong answer shows they are operating with an

incorrect theory when predicting behavior; and subjects who do know the theory about the psychological trend are more likely to answer appropriately, that is, accurately predict what they would do. This discrepancy demonstrates that our attributions and predictions are responsive to theories of folk-psychology and can change depending on learning a new theory. Therefore, we should not eliminate traditional folk-psychology on the basis that connectionist networks are also relied upon as theory when mindreading, because both tt2 and tt1 are included within the information that a person can use in order to mindread and increase accuracy of mindreading. Upon learning an improved theory, a person can train her tt1 over time to reflect that knowledge, making her tt1 more reliable. So, I propose a new TT, one that incorporates a bi-level, that is, two types of theory that are both accessed and incorporated within the phenomenon of mindreading: a folk-psychology (tt2) and a connectionist network (tt1). By incorporating both theories within a theory-based hybrid account, we will be able to address certain aspects of mindreading that may have led some to believe simulation is default over theory. This Dual Type Theory will attempt to explain the phenomenon while arguing for the connectionist network type-theory (tt1) being the default, or primary, mechanism individuals use to attribute high-level mental states to others on the grounds that this mechanism requires the least amount of cognitive resources.

Dual Type Theory. The clash between the traditional view and connectionist view of theory does not guarantee that one is incorrect about the mind. Authors argue that in many domains and for a variety of cognitive tasks people can utilize different types of processes, which are often divided between rule-based processing and associative processing (Chaiken & Trope, 1999). Under dual process theories the two paradigms could both be descriptive of the same mind, that is, the mind can be both symbolic and possess connection networks with respect to the

same task. Similarly, my view is that we can have both processes with regard to mindreading, such that a person can attribute mental states to others through use of connectionist networks, which is the default mechanism, but may also access the rule-based, symbolic theory and can use the latter to alter the former.

Dissecting theory in this way, suggesting that there are two different types, maps on to the distinction between type1/type2 processes of cognition and the associated features; therefore, I am proposing that traditional folk-psychology (tt2) resembles type 2 processing while a connectionist network theory (tt1) is a type 1 process. Better known as the distinction between system1/system 2[xiii], the two types have clusters of features that are characteristic to each, so it should be easy to identify a process as either type 1 or type 2, and classifying the mechanisms in this way might elucidate different aspects within the mindreading phenomenon like the question of primacy. I am not claiming that all dual-process theories are the same or share these exact defining features; however, many of them do, and for the sake of time I will simplify the main aspects and assumptions of the models.

Sometimes referred to as system 1 and 2, following Stanovich (1999), or old and new mind (Evans, 2010b; Stanovich, 2004), dual-process theories propose that two different processes can be used to complete the same objective with respect to a number of cognitive tasks. The two types of processes can be characterized with defining features such as, type 1 being more intuitive, not requiring working memory, and autonomous; while type 2 is reflective, requires working memory, and involves cognitive decoupling (Evans & Stanovich, 2013). Along with the defining characteristics, Evans and Stanovich (2013) propose additional, typical features of each processing mode as follows[xiv]: type 1— fast, high capacity, parallel, nonconscious, biased responses, contextualized, automatic, associative, experience-based decision making,

independent of cognitive ability; type 2— slow, capacity limited, serial, conscious, normative responses, abstract, controlled, rule-based, consequential decision making, correlated with cognitive ability.

A defining characteristic of type 2 processes, that it relies on working memory, is responsible, according to Evans (2010a), for other observed attributes. Working memory includes the operations responsible for storing information that is accessible to the mind and manipulating that information so that cognitive processes can occur. The fact that a process must use working memory consequentially underlies why that process is slow, effortful, and sequential. It has also been shown that working memory capacity, or how much information can be held accessible to the mind, varies among individuals (Daneman & Carpenter, 1980). This difference in working memory can explain the difference in ability seen in type 2 processes but not type 1. People tend to respond with the same amount of success when using type 1 cognitive modes, however there can be variety with regards to accuracy when people apply type 2 modes to cognitive tasks. Something to keep in mind, and question, is the variation in ability among mindreaders; some people are very good while others are not, and I will explain how this difference in ability is a result of varying working memory capacity, yet our default mechanism is nevertheless a type 1 process.

Evans also proposes that type 2 processing provides humans with certain cognitive abilities such as hypothetical thinking, mental simulation, and decision making (2010b). One specific cognitive ability, cognitive decoupling, is another defining characteristic of type 2 processes and relies on working memory; it is the ability to separate one's beliefs of the world from imaginary situations or being able to suppose without believing. As we will discuss in greater detail further down, simulation, which some authors argue is our default, primary,

automatic mechanism for mental attribution, possesses many of the characteristics usually ascribed to type 2 processes including using imagined situations. A potential problem for proponents of a simulation-based hybrid account or pure ST arises if we consider simulation to be a type 2 process, especially if we take 'default' to mean something like uses the least amount of resources. While there are many variations of dual-process theories each with their own definitions of two types, boundaries, and characteristics, type 2 processes tend to have a more consistent definition across theories.

Descriptions of type 1 processes, on the other hand, differ more often among theories. However, and most importantly, because a defining feature is that they are not dependent on working memory, type 1 processes are usually characterized as automatic despite varying accounts. The typical features of this group are correlated such that processes without a dependence on working memory tend to be fast, associative, and unconscious; but the correlated features are not defining of all type 1 processes. Since they do not rely on working memory, another relatively agreed upon, consequential feature is that they do not require "controlled attention" (Evans & Stanovich, 2013); so, while type 1 processing is mandatory and automatic when the triggering stimuli are encountered, type 2 processes, in contrast, require some input from high-level control systems (Stanovich, 2011). The automaticity and independence of controlled attention and working memory may provide a case for type 1 processes being the primary mode for tasks that can be completed via either process type.

The distinction between type 1 vs. type 2 processes can also be described as: unconscious vs. conscious, unintentional vs. intentional, effortless vs. effortful, and uncontrollable vs. controllable (Bargh, 1994). As previously stated and taken from the different distinctions just listed, each type has a cluster of features that are correlated with one another, such that processes

of a given type usually exhibit the cluster of features but only necessarily possess the defining characteristics. Some aspects observable in type 2 processes like attention, working memory, effort, and consciousness seem to have a connection, such that the requirement on working memory explains why these slow, conscious processes are, most importantly, effortful. The link among these features is still open to debate; but according to Baars (1997), working memory is necessary for consciousness, so to be conscious of something requires that it is within working memory, however not everything within working memory is conscious. For elements of the working memory to be conscious, there must be attention on them, so it is possible to have contents of working memory not attended to and so unconscious. Similarly, Cowan (1999) argued that "working memory includes the focus of attention, which holds the information of which the person is conscious […] However, working memory also includes activated memory outside of attention or conscious awareness" (p. 68). So when a process includes working memory it is necessarily effortful because it is utilizing limited resources; it is usually conscious because attention is focused on the element in question, but does not have to be, as consciousness is an additional but not defining feature of type 2 processes. What is important to take away is the link between effort and working memory such that a process that depends on working memory is effortful and characterized as type 2. Additionally, since we can define a process as effortful when it requires working memory resources, the least effortful procedure would then be the process that require the least amount of resources. And since type 1 does not rely on working memory, while type 2 must, we can conclude that type 1 is less effortful.

It is also worth noting that for certain cognitive functions, the mind can use either type 1 or type 2 processes to accomplish the tasks. I am proposing that mindreading is such a task, in which a person can use two types of theory, that being either traditional "folk-psychology" or her

connectionist network theory to attribute high-level mental states to others. As hinted at up to this point: folk-psychology exhibits type 2 features, and connectionist networks possess type 1 characteristics; therefore, I refer to the traditional, symbolic theory as tt2 and the associative, connectionist-network theory as tt1. As just mentioned, a person can use either of these processes to attribute mental states; however, it is less misleading to state that the tt2 can intervene with tt1 processing, such that using tt2 might provide a different attributed state than tt1. We should prefer the latter description because the default procedure is tt1, but tt2 can also be activated to reach an output and override that of tt1. For example, when asked about different heuristics that humans employ, people give intuitive responses, which are characteristic of type 1 processing but can sometimes provide inaccurate answers[xv], however type 2 processing can intervene and override the type 1 output. Usually, behavior is in line with our default processes, meaning individuals rely on type 1 processing, but intervention can occur and type 2 outputs considered when there is difficulty, novelty, and motivation. There are many motives that initiate type 2 processing, but the most prevalent is the desire for accuracy (Smith & DeCoster, 2000). Without motivation, people will use the type 1 mode, which we have identified as not relying on working memory and so effortless. But in cases of intervention comes a need for working memory resources (Evan & Stanovich, 2013). Requiring cognitive resources for working memory and attention, type 2 processes will be affected by cognitive load and susceptible to distraction and interference (Smith & DeCoster, 2000). These potential problems and limitations explain why we must be motivated to initiate type 2 processes, that is, some input from high control systems. It follows then, that in cases where either process can perform the task, our type 1 procedures, which are effortless, automatic, and do not rely on working memory, will be the primary process unless we are motivated to use and have the resources necessary for type 2 processes.

Accounts of exactly how the two processes, or cognitive types, interact can vary among the different dual-process theories. For example, Barbey and Sloman's model (2007) proposes both type 1 and type 2 processes run in parallel and resolve any conflict among the two if necessary. However, there are problems with this "parallel-competitive" view such as the waste of limited resources that are required for type 2 processes; those resources should be reserved for the most important tasks and not allocated when it is unnecessary, that is when the task can be completed using only type 1 processing. Another problem is the mismatch of time for each process. Since type 1 processes are fast and automatic type 1 would have to wait for type 2 in order to alleviate any conflict among them. (Evans & Stanovich, 2013).

Default Interventionism. A different theory of interaction and activation, default interventionism, is laid out by Evans and Stonavich (2013), in line with Kahneman & Frederick (2002), and the one I favor. The main aspect of this kind of model states that reflective logic, or reason, (type 2) can intervene and override default, intuitive responses (type 1). In a default-interventionist model, type 1 processes are assumed to reach default, automatic responses quicker than the reflective type 2 processes, but these later outputs are capable of intervening with the fast, generated ones. Type 1 processes are automatically engaged and proceed when a stimulus is encountered: however, type 2 processes, because they require limited resources, are only engaged when motivated and if the mind has enough supplies. The type 2 process uses precious resources to reach a decision but also must override the intuitive response and replace that response with the one generated by the type 2 procedure. Evan and Stanovich (2013) posit that our intuitive answers require little effort when used in novel situations, however error may occur when people lack the relevant experience. Individuals rely on type 1, even in novel situations; but error occurs because humans are cognitive misers, and thus sometimes use the

easy-to-evaluate judgment in place of a harder one even if it is less accurate. Within cognition, cognitive miser theory assumes that humans aim to conserve mental resources and so make certain attributions about the world that they are familiar with or have a bias towards. Possibly adding to the likelihood of this type of error is another factor that affects overriding intuitive responses, namely, confidence in intuitive responses. When individuals feel confident in intuitive responses, or have a meta-cognitive feeling of rightness, they are less likely to reflect on or change their answer (Thompson, Turner, & Pennycock, 2011); this rightness translates into a lack of motivation to activate the reflective type 2 mode. While accuracy is a strong motivator for activating type 2 to intervene intuitive responses in novel cases, people have a tendency to conserve resources, especially if they feel confident.

Unless there is motivation, difficulty, and novelty, a person will use type 1 processes. Type 2 processes require precious, limited working memory resources and so are only allocated when needed or motivated; therefore, people will rely on type 1 as our default cognitive mode in certain tasks since it is automatic, effortless, and does not require working memory. This preference holds true for the mindreading capacity. If in fact we do possess two different theory modes, tt1 and tt2, then tt1 will be our default procedure because it requires the least amount of resources, thus it is least effortful. And we have good reason to believe that people will conserve resources when possible, on account of the cognitive miser theory. Another reason to assert that connectionist networks, tt1, is primary comes from the definitive feature of automaticity of type 1 processes. Being a type 1 process, the connectionist network theory, tt1, is modular, fast, and unconscious, but it is also automatically engaged when in the presence of a stimulus. Thus, it is more probable that this automatic process is the one that individuals use in most cases since it is accessible in every case. Again, even though it is activated by a stimulus, it is not low-level

mindreading because it concerns peoples' desires, beliefs, etc. Because it is least effortful, tt1 is capable of activation in every situation, unlike tt2, which requires resources and motivation to proceed. Therefore, if theory is dissected in this proposed way, then tt1 is the default, relied upon in most cases, mechanism for mindreading because it requires the least amount of resources, or is effortless, and automatically proceeds when stimuli are present.

As previously stated, knowledge in tt1 exists as the strengths between nodes, but knowledge in tt2 is sentence-like information regarding the minds of others. While some connectionists claim that the existence of a system like tt1 eliminates the need for a folk-psychology like tt2, I am arguing that tt2 is necessary to help guide, increase, and correct our tt1 process. In order to alter tt1, tt2 must be accessible to consciousness and tt1 must be susceptible to supervised learning, more details to follow. For the most part, we attribute mental states to others through our tt1 unless motivated to access our tt2 knowledge. Because we rely on this intuitive response and because humans are cognitive misers, we sometimes are led astray as when we attribute things we are more familiar with instead of using limited resources. The salient features of the input might share similarities with some other input we have a great deal of experience with, and so we wrongly attribute the previously experienced output to the new input.

Understanding theory in this way, tt1 and tt2, provides this account with many upshots. The neural network of tt1 explains theory as modular and nonconceptual, such that individuals are capable of mindreading through this automatic mechanism without having to learn the concepts that are required by a traditional folk-psychology account of mindreading. At the same time, the tt2 aspect of this account explains how learning conceptual information regarding other minds can affect mindreading via motivation to access that information at the time of attribution

or through altering the tt1 connection weights through supervised learning. We also have a picture of which theory will be initiated in different situations. By appealing to the cognitive miser concept of psychology, we have an idea of what cognitive processes an individual will generally utilize, that is, the ones that require the least amount of cognitive resources. Type 1 processes are automatic and effortless, so the default mechanism will be a type 1 process. An individual will use tt1 to mindread unless the situation is difficult and novel and if the individual is motivated to use the additional resources required by tt2.

Simulation

Simulation is another proposed mechanism individuals use to attribute mental states to others. Traditional ST denies that we must use internally represented knowledge to accurately describe and predict other people's behavior. Instead, we utilize mental simulation with our own mental faculties as the model of the person to whom we are attempting to attribute mental states. One often-stated critique of traditional TT is its dependence on an elaborative folk-psychology. An implication that follows if people do invoke simulation rather than rely on some base of theories, is that the dominant explanation of cognitive capacities, internally represented knowledge, would be wrong in at least one aspect of cognition but perhaps others (Stich & Nichols, 1992). If ST is correct, then perhaps there is no knowledge structure of folk-psychology at all. Instead of relying upon theories regarding other minds, ST suggests that people use their own minds as a prototype and run a simulation to conclude and project a mental state to the target. An individual utilizes her own modules to attribute mental states that she herself would have if she simulated that situation with accurate inputs (further details concerning the mechanism are to follow.)

To some, simulation appears favorable over theory because it is not dependent on vast knowledge regarding a conceptual tacit theory of folk-psychology. It does not rely on rich information; in its place, the modules of the individual are taken "offline," so to speak, and fed pretend inputs to generate a pretend decision. For instance, to attribute a decision of future behavior of a target, one would use her own decision making module; but instead of her genuine beliefs and desires, the attributor would feed pretend desires of the target and let the module proceed to produce a pretend decision, which is then tagged and becomes a genuine belief about what the target is likely to decide (she believes that the target will do X, since she would do X if she was in that situation with the target's beliefs and desires). According to simulationists, mindreaders simulate the target's state with little or no conscious awareness. This type of mechanism is believed to produce accurate attributions since humans have the same fundamental processing features. If the mindreader puts herself in the same "starting state" and her cognitive processes execute, then there should be mental mimicry, such that her output will be the same as the target's and allow her to know what the target will do (Goldman & Jordan, 2013).

E-Imagination. In order to construct the starting states, or pretend states, a subject utilizes "E-imagination." Enactment imagination (E-imagination) is a psychological construct of mental pretense, the content of which can be "conscious or covert, voluntary or automatic" (Goldman, 2006, p. 151). The pretend states from E-imagination can be created voluntarily or automatically and can be, but do not have to be, conscious. In some cases, a subject may engage consciously to recreate a perceptual experience through E-imagination, like trying to remember where her dresser is in relation to her bed. This recreated state would be conscious and voluntary. Another description of E-imagination given by Goldman and Jordan (2013) is the faculty that constructs a state like the specific state one wishes to be in. If a person wishes to be in mental

state M, she can imagine being in an M-like state, which is functionally very similar to M, and when either are fed into cognitive mechanisms similar outputs are computed. In the case of predicting what a person will decide to do, the attributor imagines similar beliefs and desires of the target to be fed into her own decision making module. Since her decision-making module is similar to the target's, the attributor's module should produce a similar or identical output. The offline, pretend output is then sent to the attributor's prediction module resulting in a prediction of what the target will do. Again, this process occurs without relying on information or theories about how other minds operate (e.g. make decisions); the attributor comes to an attributable state through simulating what she would do (what decision she would make) if she held starting states (e.g. beliefs and desires) similar to the target.

Let's revisit the example of Donna and the donuts, but this time proceed according to ST. Donna desires donuts and believes that DonutDisturb has the best donuts; so, which shop will she go to? In order to compute what Donna will decide, we need to imagine what we would do by simulating her mental state. An individual would need to create pretend states (from E-imagination) that match the states of the target, and then she would feed these pretend states into her own decision-making module, which is being run offline, to generate a decision. In the simulation mechanism, the inputs are the pretend states and the output is the state which will be projected and attributed to the target. For Donna's case, the mindreader would construct a pretend desire for donuts and a pretend belief that DonutDisturb is the best donut place; the generated decision to go to DonutDisturb is still a pretend output. Since the modules during simulation are taken offline, the outputs are still not genuine (meaning not belonging to the mindreader)[xvi]; however, at this point the simulation process ends and the output is then

attributed to the target and becomes a genuine metarepresentational belief of the mindreader (i.e. the belief that "Donna will go to DonutDisturb"). See Figure 2.

Simulationists claim that ST is simpler than TT, however that is not completely obvious when we identify the two components necessary for mindreading: the database of how people behave and the mechanism which extracts that information (Stich & Nichols, 1992). Earlier in the discussion about traditional TT, I mentioned that inference mechanisms are used to extract information from other folk-theories, like folk-physics, so we can use this same tool for our folk-psychology; however, TT relies on a database of encoded information of how other minds work. So, while the mechanism for TT is already given, the database is not[xvii]. ST, on the other hand, has the database "for free" since it uses the attributor's own mind as the model; but the mechanism of extracting the relevant information is not as freely given as that of TT. For ST, the module must be taken offline, pretend states are generated and fed into that module, and the outputs have to be transformed from pretend outputs into metarepresentational beliefs concerning the target (that is the attributor believes that the target believes X, or will do Y)[xviii].

Although simulation does not rely on some vast information or folk-psychology, it does involve certain key skills in order to operate. Imagination, or as Goldman (2006) called it "E-imagination," self-reflection, inhibition, and characterization of the output are all distinct features and necessary elements of simulation. Goldman (2006) proposed that self-reflection, or introspection, is the tool within the mindreading phenomenon that not only initiates simulation, but also is necessary to inhibit one's genuine states and to characterize and attribute the generated outputs from simulation. Mitchell, Banaji, and Macrae (2005) demonstrate that self-reflection is a natural subactivity of third person mindreading by citing studies, in which the medial prefrontal cortex (MPFC) was activated during tasks requiring introspection and also

when completing a study concerning third-person attributions; the results of which were consistent with earlier findings that the MPFC is activated during third-person mindreading. Therefore, people do engage in self-reflection during many higher-level mindreading tasks. Goldman states that self-reflection is not as explicitly or obviously part of TT, and so the empirical evidence of activated or engaged self-reflection is reason to support a ST-based account (2006)[xix].

To engage simulation, an individual has to judge herself as similar to the target; and this occurs through self-reflection. If she does not consider the target to be like her, then she will not attempt to simulate that mental state because being so dissimilar might mean that she would not be able to reach an analogous pretend output. The judgement regarding similarity is not always conscious, but it is necessary to start the simulative mechanism. Perhaps the failure to simulate based on judged dissimilarity is evidence that simulation requires effort, such that individuals only simulate when they believe it can provide accurate attributions (they simulate to explain the behavior of a person of the same race but not a member of some outgroup, despite the behavior, or input, being the same for both individuals).

Inhibition. In order to simulate someone else's mental state, a person must inhibit her own genuine beliefs etc. and not include them in the simulation. Therefore, reliable simulation involves a person monitoring her own genuine states and inhibiting or quarantining them. Self-reflection is the tool used to identify and quarantine one's genuine states, so that they do not contaminate the simulated output. Without this step, quarantine failure can occur, such that the simulated output would not reflect the belief of the target because the subject's genuine states (beliefs, feelings, etc.) are affecting the simulation. The interference of genuine states can lead to error unless the mindreader's own mental states are inhibited. Inhibition in this context is closer

in line with self-control, rather than inhibition in the neural sense. In the neural sense, the activation of a neuron is inhibited; but in mindreading, inhibition refers to genuine states to be tagged, separated from pretend states, and disregarded during the simulation via introspection. Inhibition, or quarantine, is extremely important as a mindreader will always have her own desires, beliefs, and intentions alongside the pretend ones; her genuine states must be segregated from the pretend ones, an activity that may not be trivial in producing an accurate attribution (Goldman & Jordan, 2013). The dependence on inhibition in mindreading and the difficulty associated with it leads some, like Goldman and Jordan (2013), to believe that simulation is the best explanation for egocentric bias, which is a common error in mindreading. The phenomenon of egocentric bias occurs when a person incorrectly attributes her own mental states to others. This phenomenon is potentially problematic for a TT model because it seems to suggest that attributing mental states to others relies on access to one's own or at least incorporates it, which is not consistent with traditional TT (Nichols, *forthcoming*). Because this type of bias is so prevalent in mindreading and can be explained by appealing to our genuine states that are not quarantined in simulation, some take the phenomenon of egocentric bias as evidence for simulation as the default mechanism in attributing mental states to others. This phenomenon will be addressed and explained via the Dual Type Theory Account later.

How do we go about inhibiting these genuine mental states? Do we really rely on a process that is so difficult, such that mindreaders often fail to inhibit their own states and incorrectly attribute contaminated outputs? Proponents of simulation argue that egocentric bias is evidence for simulation being our default mechanism; however, appealing to the difficulty of the process seems inconsistent with arguing that it is default[xx]. If quarantining genuine states in mindreading involves stopping those states from being fed into the offline module, then that type

of control probably requires the same amount of effort as stopping certain inputs from feeding into modules when not offline. For example, when deciding how to respond to someone who has treated her wrongly, but wanting to remain in control and not react out of extreme anger or frustration, does a person inhibit the output of that module given the input of extreme anger or does she stop the extreme anger from being fed into that module? It is not obvious, but I think the latter approach is more realistic with respect to efficiency of being able to control one's behavior and emotion regulation. It is easier to stop the consequent behavior that an emotion might elicit when that emotion arises, not after it has been processed and a resulting behavior has been decided.

If emotion regulation, or other instances of self-moderation or self-control, rely on the same type of inhibition that simulation postulates[xxi], then we can look at instances of the former and its relationship with cognitive resources to provide insight into whether a mechanism that incorporates this type of inhibition is going to be default. A more in-depth discussion about egocentric bias, defaultness, and the implications on mindreading mechanisms is to come when we discuss hybrid accounts; however, I would like to quickly mention some empirical studies that have documented cognitive resources affecting one's ability to self-moderate concerning emotion regulation. If emotion regulation requires cognitive resources, and if emotion regulation involves the same type of inhibition as simulation, then simulation is also taxing on cognitive resources.

Grillon, Quispe-Escudero, Mathur, and Ernst (2015) tested whether individuals depleted of cognitive resources would be worse at regulating their emotional responses. They found that the group of participants who performed a difficult task were not able to decrease their emotional response like the participants who did not complete a hard task, despite both groups having

similar startle responses. Both groups experienced emotion reactivity, however the group that was mentally fatigued, and therefore depleted of cognitive resources, was not able to downregulate that emotional response. One possible explanation is that they were unable to inhibit this emotion from being an input within certain modules because the resources needed for inhibition were depleted.

Johns, Inzlicht, and Shmader (2008) propose the difficulty associated with emotional regulation as the explanation for stereotype threat. When aware of a certain stereotype, those that fall into that category often perform according to the stereotype, which is negative. The process responsible for decreased performance as seen in stereotype threat instances can be explained by a subject's attempt to control her emotions. The regulation of emotions depletes cognitive resources, resulting in worse performance. However, if participants are given coping mechanisms to control emotion, they do better on the tasks.

In another study, participants specifically instructed to regulate emotion (hide frustration) performed worse; the increase in errors suggests that emotion regulation depletes attentional resources as well (Goldber & Grandey, 2007). Testing the other direction of this relationship, Schweizer, Grahn, Hampshire, Mobbs, and Dalgleish (2013) show that easy working memory training can lead to improvements in emotion regulation, such that participants who received the training reported being better at reducing negative emotional responses. This result comes from training working memory and increasing the ability to perform exhausting or difficult tasks over time. The better the individuals are at using working memory resources effectively, the better they are at emotion regulation.

Characterization. While simulationists attack TT for being complex, simulation is not as simple as it seems. Not only does inhibition, which is necessary for the mechanism to

appropriately mimic the mind state of another, deplete cognitive abilities, another potential problem that arises concerns characterizing the output of the simulation from a pretend belief to a metabelief attributable to the target. In simulation, the offline module provides a pretend output that the subject must attribute to the target by tagging it as a metarepresentational belief (a belief about what the target believes). How does this occur? Goldman discusses how an individual can go from a pretend state x to a genuine belief that T believes x, a metarepresentation that incorporates attitude type and content, via self-reflection. He suggests that self-reflection, or introspection, is capable of classifying the pretend, simulated output by tagging it as an attitude of the target with certain content. And as previously stated, self-reflection is necessary in ST and proven to play a role in mindreading tasks; it is less obviously included in TT. Therefore, Goldman takes the role of self-reflection in this instance as more evidence for a simulation-based hybrid account. The empirical evidence of self-reflection in mindreading tasks is explained by the role it plays in simulation to inhibit genuine states and to tag outputs. However, I argue that there is also a role of self-reflection in theory-based mindreading. I will provide an alternative explanation to address this concern[xxii].

If mindreading occurs via simulation, then the accuracy of the attribution relies on the starting states, or inputs, of the simulation. Given that humans have the same processing, if we use our own modules as a prototype and supply starting states believed to be similar to, or the same as, the target, then the module will produce an output that is similar to, or the same as, that of the target. Our similar psychology ensures accurate attribution by mimicking the mental state. However, if the mindreader does not imagine the appropriate starting state or does not possess the right information, then the output will most likely be incorrect (Goldman & Jordan, 2013). If a module is fed incorrect inputs, then the output will not be like the target's. In general, E-

imagination accuracy depends on relevant knowledge, such as memory, to create accurate inputs for simulation. For example, trying to simulate an image of the Titanic without having seen it or having detailed descriptions would generate a less reliable, or accurate, visual imagery than a simulation that accesses past experiences of seeing pictures or having seen the ship in real life. In the traditional ST-TT debate, the question arose whether visual imagery, like that of the Titanic, which relies on information, is simulation or theory since the "generation involves recourse to information stored in memory" (Goldman, 2006, p. 150). However, according to Goldman, reliance on information in memory does not make a process non-simulative, as long as the state is produced top-down for the purpose of replicating a naturally produced state (the imagined visual image of the Titanic replicates the state of seeing the Titanic).

In addition to lack of relevant knowledge limiting the construction of the pretend state, another problem is the necessity to inhibit genuine states, as states earlier. The genuine states must be separated and inhibited, otherwise the computed output will be contaminated and the simulation not accurate. To avoid this interference, the genuine states are "quarantined" so that they do not penetrate the imagined or simulated tracking process (Goldman & Jordan, 2013). It is effortful, as seen in comparable instances of inhibition like emotion regulation. So, if humans use simulation, and therefore inhibition, to attribute high-level mental states to others, then we are using a model that depletes cognitive resources. Conversely, if a person is depleted of cognitive resources, she will be less accurate since she does not have the means needed to quarantine her own states.

But can simulation really be our default mindreading mechanism with so much reliance on limited cognitive resources? As I have identified 'default,' and when considering the Dual-Type Theory and default-interventionism, there is reason to believe that type 1 processes will be

default mechanisms for a given task, but type 2 theory can override or intervene when the situation is novel, difficult, or when there is motivation (like the greatest motive: a desire for accuracy). Without motivation, people will use the most effortless mode available. Such a process would be one that possesses type 1 features: intuitive, automatic, and does not require working memory; thus, if a system or process does require working memory, then it is not type 1, but type 2. Simulation might be a type 2 process, since it is affected by cognitive load or resource depletion, and it possesses other defining features of type 2 mechanisms: reflective, requires working memory, and involves cognitive decoupling (Evans & Stanovich, 2013). According to Evans, (2010b), type 2 processing is also responsible for some cognitive functions like hypothetical thinking, mental simulation, and cognitive decoupling. Cognitive decoupling is the ability to separate supposition and belief, such that a person can imagine situations without believing them. Simulation in high-level mindreading, the process of imagining certain states to create pretend states that can be fed into one's offline modules in order to determine and attribute high-level mental states to others, seems to include cognitive decoupling. It also relies on working memory and cognitive resources, since it can be affected by cognitive depletion. If simulation is a type 2 process, then maybe this is more evidence that it is not our default mindreading mechanism, especially if we can provide an alternative method that is reliable but not as effortful. Simulation is fast and unconscious, so perhaps it does not fit into type 1 or type 2 but is a hybrid of some kind; and while tt2 is not fast and unconscious, like simulation, tt1 is less effortful. Because authors have not divided theory in this way, this type of reasoning was not available in previous discussions about primacy.

In the traditional ST- TT debate, the theory in question was sentence-like, rule-based, internally represented knowledge regarding the minds of others; and it is also effortful since it is

deliberate and requires accessing that information in memory systems (Smith & DeCoster, 2000). However, by including a connectionist network (tt1) in addition to the traditional idea of theory (tt2), we have an alternative method of attributing mental states to others that does not rely on an internal database of sentence-like folk-psychology, is less effortful because it does not use limited working memory resources, and can also explain some of the empirical evidence of mindreading. By appealing to connectionist networks, tt1, we still have knowledge representing the world, but it is not like that of tt2. Instead, tt1 is synchronically state-nonconceptual and the strengths among the nodes reflect the statistical probabilities of regularities within the environment. While I will argue that our conceptual tt2 knowledge can train or adjust the connection strengths in our tt1 to reflect the knowledge found in tt2, making that network diachronically conceptual, the synchronically nonconceptuality of tt1 provides the networks with automaticity and effortlessness. Therefore, in cases where it can provide accurate associations of the environment, tt1 will be the default mechanism[xxiii].

Explanations of a Hybrid

Traditionally authors argued for one mindreading method exclusively over another, that is that an individual performs all mindreading only through theory or only through simulation. However, mindreading is a complex construct that utilizes different approaches depending on the type of mindreading task at hand, motivation, and cognitive resources available; therefore, there has been increasing support for a hybrid account that includes both theory and simulation. Within these hybrid accounts, one mechanism is still given primacy over the other, or considered our default process behind attributing mental states to others.

The variety of mindreading tasks leads authors to propose hybrid accounts like that of Stich and Nichols, who favor theory over simulation, and Goldman, who holds a simulation-

based account. I will not focus on defending the existence of a mindreading capacity that incorporates both approaches, but I will quickly mention some instances that establish both mechanisms are used in mindreading. Stich and Nichols (2003) provide cases for both simulation and theory mindreading to argue for a hybrid account, since the examples include obvious instances of one process being able to provide a conclusion when the other cannot.

A case of mindreading which is most easily understood by appealing to simulation is inference prediction; people are capable of predicting inferences of targets, even when they are nondemonstrative (Stich & Nichols, 2003). TT advocates have not tried to explain this phenomenon while simulation provides a simple account that addresses the accuracy of inference prediction. A way that TT could try to explain this is to suggest we have a theory about how others will make inferences. However, it is simpler to suggest that we rely on what our own inference would be and attribute that to the target. Not only is it simpler to appeal to simulation, it also explains the accuracy of this skill which encompasses instances that we have no experience with. Individuals can accurately predict inferences even in cases which are different from things they have likely encountered; so, it is more probable to propose that we rely on what our own inferences would be (simulation) and attribute those to others than to suggest we have somehow acquired information, or a theory, about how people will reason (which is farfetched when considering that we are good at predicting inferences even when we do not have experience). The accuracy aspect of this phenomenon is reason to believe that we are using our own inference mechanism to simulate what our inference would be.

Desire attribution, on the other hand, cannot be explained by appealing to simulation and instead demonstrates an instance of mindreading that must be subserved by theory (Stich & Nichols, 2003). When an individual attributes the desire that caused the target to behave in such

a way, she uses theory. Goldman (2006) agrees that in this kind of retrodictive mindreading, it would be impossible to use simulation alone because modules are unidirectional, and therefore cannot take the outputs and simulate what belief or desire brought about the behavior or belief in question. He proposes it is more likely that individuals theorize and conclude certain inputs to then test in a simulation, similar to hypotheses testing. In order to explain the behavior she just observed, an individual will rely on theories to gain insight into potential reasons why a person said or did a specific thing. She then uses the insight from the theories as inputs for a simulation and compares the simulated output with the observed behavior. However, a problem with this analysis-by-synthesis account is that it generates too many candidates to be tested. The underdetermination that arises with the simulations from theorized inputs, in addition to the systematic inaccuracy of desire attribution, suggests that it is not simulation which leads to the attribution (Stich & Nichols, 2003). Desire attribution may just be reliant on theory and not the subsequent testing via simulation. Regardless, desire attribution is evidence of mindreading via theory.

Mindreading incorporates both methods, but again, the principle concern of this discussion is which mechanism is our default. By saying one process is default usually means it is the automatic, standard, or "most basic and spontaneous method" (Goldman & Shanton, *forthcoming*). Goldman mentions some potentially phylogenetic[xxiv] and ontological[xxv] reasons, but ultimately his empirical reasoning (such as egocentric bias) supports the claim that we will use simulation as the default mechanism in a hybrid account. I disagree and propose theory is our default, so I will focus on his empirical reasoning and provide alternative interpretations that attempt to explain certain phenomena through a Dual Type Theory Account. I think appealing to the least effortful mechanism is a better determining factor of the default process. As explained

through the psychological aspect of cognitive misers, without motivation people will use the most effortless mode available; and this phenomenon is evident in other aspects of cognition.

Certain psychological tendencies, like the 'representative heuristic,' demonstrate that individuals respond intuitively because it requires the least amount of effort (Goldstein & Gigerenzer, 2002; Hilbig & Pohl, 2008). The representative heuristic states that a person relies on past experiences to guide decision making, so when a person is choosing between two objects, and she has experience with one over the other, she will choose the one she recognizes. This is different than the "mere exposure effect," which suggests that people prefer things, with which they are familiar. The representative heuristic is more of a shortcut for individuals to use representation they have experience with to guide behavior or decisions in novel situations. We can apply the same logic of the representative heuristic to other mental processes like mindreading. When confronted with situations we are familiar with or have similarities to past experiences, the cognitive system will come to a decision based on convenience and speed, thus using the least amount of cognitive resources. With the descriptions of type 1/type 2 processing, individuals will choose type 1 processes, which are automatic and do not rely on working memory resources. The bat and ball problem[xxvi] is another example of reliance on type 1 processes, according to Kahneman (2011), when people place too much faith in intuitive responses because of their aversion to using cognitive effort.

While I will make an argument for a theory-based account of mindreading that appeals to the amount of effort of each method as evidence for tt1 being default, first I will present some of Goldman's empirical reasons for the primacy of simulation. To support his claims, he references the pattern of error known as egocentric bias as evidence that simulation is primary since this phenomenon can be explained through quarantine failure during simulation. He also cites studies

regarding self-reflection used during third-person mindreading tasks in support for simulation-based mindreading. Self-reflection is known to be used in the simulation process, but it is not obviously included in a theory account (Goldman, 2006). However, we can consider interpretations that explain these phenomena as compatible with a Dual Type Theory Account.

Egocentric Bias

The seemingly most convincing empirical evidence for simulationists and problematic for theory theorists is egocentric bias, which is explainable through simulation, but not as obviously through theory. Egocentric bias, a common bias in mindreading, occurs when a person attributes her own mental states to others. It appears to suggest that we rely on access to our own states when attributing mental states to others, in which case mindreading would be based on simulation rather than some tacit theory (Goldman & Shanton, *forthcoming*). Because egocentric bias is so prevalent in instances of mindreading, if it does imply simulation, then this is reason to believe that simulation is the default mechanism in a hybrid account of mindreading. Therefore, an account based on theory must address this phenomenon.

According to simulation-based accounts of mindreading, egocentric bias is explained through quarantine-failure, or the failure of inhibiting one's own genuine states when simulating the mental states of others through imagination. ST proposes that in order to mindread higher-level states of a target, the mindreader uses her own modules as the prototype of the target's mind and feeds pretend states, that resemble states of the target, into her offline modules to reach a pretend output, which is then characterized as a metarepresentational state attributed to the target. The accuracy of the output depends on the resemblance of the starting states, which the mindreader creates through E-imagination, to the actual starting states of the target. Since humans share similar mental processing, the attribution should be accurate if the inputs are

similar to the target's starting inputs. However, another aspect affects the correctness of the simulation process. As discussed earlier in the simulation section, a mindreader must separate her own genuine states from the pretend ones fed into her offline module. Although she creates pretend beliefs that reflect what the target beliefs, she still possesses her own beliefs, desires, etc. If her own states are included in the simulation, then the output will not reflect the target, who may not share those beliefs, which are genuine to the mindreader since the inputs would differ between the subject and target. Therefore, inhibition, or quarantine, becomes extremely important for accurate attribution by excluding genuine beliefs from the simulation.

If ST is correct, and we construct pretend states to be fed into our offline modules, then the pretend states will contain similar elements to our genuine states; and this similarity will make confusing them likely to occur (Goldman & Shanton, *forthcoming*). If our modules confuse these two bodies of states, then our genuine states could replace the pretend ones in the simulation process; in which case, it would produce an egocentric output. Therefore, an individual must inhibit genuine states, that is separate and tag them to be excluded from the simulation. As discussed earlier, inhibition is difficult and requires cognitive resources; thus, the difficulty would explain the pervasive failure and consequent egocentric bias. As seen in instances of self-regulation and emotion-regulation, which also incorporate inhibition, the ability to inhibit is affected by the cognitive resources available (see discussion on simulation stated earlier). Since it is affected by cognitive depletion, inhibition is effortful; and so, simulation (or accurate simulation) is also effortful. While the difficulty to inhibit is one approach to explain egocentric bias, we may also appeal to theory to understand the common error without the demand on cognitive resources.

However, Goldman & Jordan (2013) claim that TT does not readily explain this phenomenon, since states associated with it are not as easily confused with one's genuine states, whereas the pretend states of ST closely resemble the same type of states as a mindreader's genuine ones. Throughout mindreading literature there are many types of egocentric biases, like false-belief attribution, in which individuals allow their own knowledge to affect the attributions to their targets. With a hybrid account of mindreading, I propose that some instances of egocentric bias are a result of simulation and the difficulty of inhibition that is associated with it. Some instances of mindreading are a result of simulation, yet theory is our default mechanism. While some egocentric attributions are from a failure to inhibit genuine beliefs during simulation, other egocentric outputs could be the result of our theories used in mindreading. Therefore, egocentric bias is compatible with a theory-based account. I think that we can develop egocentric theories, that, when incorporated into our tt1, ultimately affect how we attribute mental states to others. We can offer an explanation via theory, specifically the Dual Type Theory Account, to explain many instances of this pervasive egocentric error in order to hold tt1 as the default mechanism of mindreading.

An account arguing that tt1, or the connectionist network theory, is the default mindreading process, must account for egocentric error and offer an interpretation of the phenomenon while appealing to the encoded theory in tt1. As Nichols (*forthcoming*) stated, there is reason to consider egocentric attribution as the default response that people use unless there is overriding information that provides more clarification; additionally, this bias increases if people are under time pressure (Epley, Keysar, Van Boven, & Gilovich, 2004). If our intuitive response is egocentric, then we seem to be using our own minds to understand the minds of others; so, our knowledge of other minds depends on our own experiences (Nichols, *forthcoming*). But can we

explain this egocentric knowledge in terms of a Dual Type Theory Account? If our intuitive responses are a result of tt1, then tt1 must be egocentric. The egocentric theory encoded in tt1 could be reflecting an egocentric tt2, if our two systems interact in that way, such that tt2 can train tt1. What I am suggesting is that our knowledge regarding other minds depends on our own mind and experiences, which then are generalized and added into our folk-psychology. We use our own case to make inferences about all other minds. If we do generalize our experiences in this way, then our folk-psychology theory will be inherently egocentrically biased. Although this seems problematic to make generalizations of folk-psychology based on one case, the egocentric strategy is highly reliable (Nichols, *forthcoming*). Especially if human psychology is similar, a folk-psychology incorporating generalizations of our own mental processes will be highly accurate in understanding other minds. However, we need an account of an egocentric tt1, not just folk-psychology (tt2).

Acquisition

To discuss egocentric tendencies on the Dual Type Theory Account, we need to first understand acquisition in both theories and the connection among them. As mentioned before, tt1 is the connectionist network representing knowledge regarding other minds via the strengths of connections among the nodes within the network. I am proposing that connectionist networks can learn in a bottom-up way, but also in a theoretically-driven way. These networks are synchronically state-nonconceptual, and therefore do not necessarily depend on concepts; in which case, they reflect the statistical probabilities and regularities experienced in one's environment. However, they are susceptible to supervised learning and so can alter the weights within the system with influence from tt2, conceptual knowledge of folk-psychology. A more in depth discussion of this process is to follow. Although this process is slow and effortful since it

requires consciousness, it is possible for type 2 systems to train type 1 via consolidation, which includes repeated presentations from the type 2 to type 1 processes (McClelland, McNaughton, & O'Reilly, 1995).

Tt1, being modular, implies the innate capacity that humans have to mindread, however the reliability of those connectionist networks depends on the adjustment of the weights among nodes in response to cues in the environment and successful responses to that environment. If new situations are encountered, ones that include stimuli that are not inputs in a current connectionist network, then the connectionist network will not be reliable. So this explains why children's theory of mind is not well developed, because they lack the experience with other minds to have a reliable tt1. Instead, a child uses similar inputs that sometimes produce inaccurate outputs. However, as the child learns new concepts and increases her tt2, she can use this information to train up her tt1 to respond appropriately. This requires allocating attention and effort, which increases in ease as a child becomes older. Tt1 is responsive to environmental cues but also to tt2; so, inaccurate attributions could be the result of relying on similar, but incorrect, inputs of a network, or they could be stem from using a network that is guided by an inaccurate tt2. As our tt2 knowledge increases and our allocation of attention improves, we can train our tt1 to reflect accurate knowledge regarding other minds.

In regards to egocentric bias, this is not the hard-hitting case for a simulation based hybrid account that Goldman thinks it is. Our default tt1 can be egocentric if it is trained by supervised learning via tt2, which exhibits egocentric theories. But how does our tt2 gain the knowledge it has about the psychological processes of other minds? Folk-psychology, is the information regarding the minds of others, not our own psychological occurrences; however, we use our own psychological patterns and our own experiences to generalize about the habits and

thoughts of other people, or at least this is where our folk-psychology theory of mind begins. Is there a mechanism available to us that is capable of such a task? As seen in Goldman's account of projection from the simulation mechanism, introspection tags our outputs as a metarepresentational belief to be attributed to the subject in question. It is just as probable, then, that outputs can also increase our tt2 in the same way, that is the output is characterized and sent to some other module, like an inference module, to make inferences about other minds. In the case of simulation, our belief generation module and others run offline; but if introspection is capable of tagging these offline conclusions, it can also tag the ones that truly belong to us—our genuine feelings, reasons, and experiences. If self-reflection can tag outputs to be classified in a different way, then we could use our own experiences to generalize about the minds of others. Those experiences, the outputs of our modules, become included in our folk-psychology. In this case, if our theory of other minds can come from the generalization of our own experiences, then this knowledge will automatically possess an egocentric foundation. Therefore, a tt1 that is trained by such a tt2 will also reflect an egocentric bias.

Now there are cases in which we do not have experiences with that given situation and do not have the tt1 to reliably attribute mental states to others. In this case, we are motivated to use additional resources for a more accurate attribution. We can run a simulation to determine how we would feel and then attribute that state to the person in question. But once that experience and tagging is successful, that information is now added to tt2 and up for availability to train our tt1. The outputs of simulation count as experiences to be generalized and added to the tt2. Once they are included in tt2, that knowledge can train out tt1 through supervised learning. I am suggesting, though, that our primary mechanism is tt1, but if we do not possess the correct network for a given circumstance we can resort to simulation to gain that information. As a child's theory of

mind is developing, she does not have reliable networks nor a vast conceptual knowledge regarding other minds, instead she relies on the generalizations from her own experiences, or in cases where she does not have experience, she will simulate, come to some conclusion about how she would feel, generalize that output to other minds, and add that knowledge into her tt2. The innateness of tt1 might explain how humans have one folk-psychology for everyone. As a child increases her knowledge of other minds and learns new concepts, she can begin to expand her tt2, which can then influence her tt1. This Dual Type Theory also explains how we can have intuitive egocentric responses (tt1), but more information can alter the output (tt2 intervenes). Much like that of folk-grammar, our learned conceptual knowledge can help aid our executing the task at hand, but the default procedure, the connectionist network, is still responsive to stimuli and experiences (such as hearing certain grammar and learning from that, without having accessible theory). This tt1 is modular in that it is innately predisposed, but our conceptual learning can override that disposition and then modularize it, such that the tt1 is still automatic but diachronically conceptual.

Self-Reflection

This account of Dual Type Theory as it explains egocentric bias via the acquisition of tt1 and tt2, additionally addresses another empirical aspect of mindreading. The studies which demonstrate self-reflection as occurring during instances of mindreading has led some to support a simulation-based hybrid since ST incorporates self-reflection or introspection as a necessary component while traditional TT does not inherently suggest it (Goldman, 2006). Since self-reflection has been found to be activated in instances of mindreading, a good account of mindreading should have an explanation of how self-reflection can come into play, which traditionally has been more obvious for ST than TT. In simulation, introspection is required to

tag and quarantine one's genuine states so that the offline module computes only the pretend states believed to belong to the target. Introspection is also the tool that tags and categorizes the output of the simulation as a metarepresentational state attributable to the target. While observed activation of areas believed to be connected to self-reflection may be caused by simulation during mindreading, this is not the only possible explanation, and therefore does not guarantee that simulation is our default mechanism. Although self-reflection is not necessary for the automatic activation of tt1, perhaps it is the tool that attributes the output of the connectionist network as a metarepresentational state to the target. Similar to its function in simulation, self-reflection is capable of tagging these states and categorizing them. If self-reflection is the tool tagging and attributing the output states in simulation cases, then similarly the same tool is used in attributing the outputs from the theory cases. Tt1 produces an output, but it must be attributed to belonging to the target by some means. Self-reflection appears to serve this function in simulation, and so could also be that tool in mindreading via theory. If self-reflection is involved in this way, then we would see activation in cases of mindreading via theory as well. Therefore, the activation of self-reflection regions in the brain is compatible with theory-based accounts.

Additionally, self-reflection occurs with respect to type 1 processing when an individual reflects on the success and failure of her processing[xxvii]. This type of introspection can alter the weights among the nodes of the connectionist network; a process necessary for reliability. Accurate attributions will then increase in strength while incorrect ones will cause the connection to weaken. Also, self-reflection adds symbolic knowledge to her rule-based theory (tt2). If self-reflection can tag pretend outputs, then it can tag and characterize genuine ones too, such that it could generalize them into our folk-psychology. We can use self-reflection to add to our tt2 by characterizing some output or genuine state to be representing all humans not just our own case.

This type of addition to tt2 is responsible for the egocentric bias and facilitated through introspection. So evidence of self-reflection during mindreading could be cases of reflecting on the success of the connectionist network for the sake of improving reliability of tt1 or tt2, it could be the result of tagging genuine outputs as new theories within tt2, or it could be tagging the outputs of the connectionist network, tt1, as metarepresentational.

Other Phenomena

Some additional reasons for thinking that our mindreading is mostly theory-based include the cognitively penetrable aspect of attributions, such that our learned theories can greatly change which attributions we make, in addition to the roadblocks and potential consequences associated with simulation. Our attributions have been shown to be influenced by new theories of the attributor (Stich & Nichols, 1992), as discussed earlier; but this is not compatible with simulation. If we did rely on simulation to make most of our attributions about the mental states of others, then learning a new theory would not change the output considering our offline module is responsible for producing the pretend state attributed to the target given the pretend starting state believed to belong to the target. So, a new theory affecting our attributions is evidence that simulation is not responsible.

There are also instances in which we cannot or are not motivated to simulate, and this lack of automaticity might suggest that it is not our default process. The simulation process is triggered by a "like-me premise;" and the fact that simulation relies on the perceived similarity between the attributor and target suggests to me that some theory is being activated, and as a result that activation triggers simulation to occur. The theory behind the like-me premise guides simulation. Without this theory, simulation would not occur without being motivated (e.g. the subject is reminded or requested to do so). This phenomenon of failing to simulate based on the

lack of similarity is observable in studies like that of by Kaufman and Libby (2012), which demonstrates that perspective-taking increases when the subject believes the target to be like her as opposed to a member of an outgroup. The fact that we do not engage in simulation with people we think are not like us, may be evidence that we do not want to waste resources when the output of our simulation will not be relevant, since the target differs from us. If this is true, then refraining from simulation in cases deemed too different could show that simulation requires precious resources, and therefore is effortful. If we have tt1, which is not effortful, and that mechanism has the appropriate inputs and will be reliable in the current situation (meaning we are not motivated by accuracy to use limited resources), then we should hold the view that tt1, not simulation, is our default mechanism, that is, the one that is relied upon most since it requires less resources, given the cognitive miser theory, which states that individuals will conserve resources. Additionally, if we do not simulate, then it is not a type 1 process, given that type 1 processes automatically proceed when the stimulus is present.

Current emotion also has a strong influence over simulation, sometimes making it nearly impossible, which might give support for why it is not the primary mechanism since there can be many instances, in which simulation is not achievable. It could be that the emotion has such intensity that it cannot be inhibited, as when we find out that a parent has died. No matter how hard we try, it seems like our own situation, in such a case, cannot be ignored; simulating a very different mental state, like the joy from winning the lottery, seems impossible. Or it could be that we are depleted of resources from attempting to regulate our emotion, and there are not available resources left for simulation. However, even in these cases we can still attribute mental states to other people. Since simulation relies on the inhibition of one's genuine states, and since inhibition requires cognitive resources, a lack of these resources or a high demand of these

resources might affect whether a person indeed simulates. The person still attributes mental states; and this ability to attribute, even in cases where inhibition of genuine resources is not likely, is strong evidence that mindreading is executed via theory. Someone who favors simulation might respond by stating that while simulation is impossible in those cases, for the most part, it is default. However, these cases show simulation requires resources, and the demand of resources makes it unlikely that simulation is default, especially when another reliable process is available and less effortful.

Another reason that might suggest that our primary process for mindreading is tt1 appeals to the potential negative consequences of simulation. Goldman claims that quarantine failure is possible when the pretend states and genuine states, which are similar, are confused during simulation and genuine states are incorrectly fed into our offline modules. It is just as likely that quarantine failure can also occur on the other side of simulation, that is the pretend output could be confused and attributed to the mindreader as a genuine state. Theory on the other hand does not predict this possibility. Quarantine failure of pretend outputs could be problematic, especially if simulating negative mental states; thus resulting in submitting oneself to those negative mental states, which have potentially negative side effects such as lower academic performance, worse memory, physical illness, and affecting subsequent thoughts. Xie and Zhang (2016) found that positive emotions increased accuracy of face discrimination while negative affections impaired performance. The risk of quarantine failure happening on the output side of simulation would allow for such negative effects to occur to the attributor. If quarantine failure is as prevalent as simulation-based accounts concede, such that, it explains our pervasive egocentric bias, then the same failure to distinguish one type of state from the other (genuine from pretend, or vice versa) predicts that simulation of high-level mental states could affect the genuine states of the

mindreader. If quarantine failure occurs so often in cases of egocentric bias, it could be expected to occur with outputs as well. If quarantine failure occurred with outputs, we would expect to observe effects like those just mentioned. However, attributors do not seem to confuse attributed beliefs or desires with their genuine ones. Quarantine failure could occur if simulating, so we would expect to see some effects, but we do not. Additionally, these effects could be unproductive for the subject. Since humans do not face this same threat of quarantine failure with the outputs, perhaps we do not use simulation as default, but rely on theory.

Dual Type Theory Account

I have argued bits and pieces, here and there, throughout this paper; however, I would like to bring some of it together to present how the tt1 of a Dual Type Theory Account could be the default mechanism one uses in attributing higher-level mental states to others. I have shown that Goldman's arguments for the primacy of simulation are not as convincing as he hopes and a Dual Type Theory can equally explain the phenomena. Therefore, I think we should consider again how cognitive systems select mechanisms in a given situation. In regards to what function is default, we should accept the idea, "using the least amount of cognitive resources" as a good marker of which mechanism the cognitive system will select in most cases. Because effort can be thought of as using cognitive resources, the default process will be the least effortful. If a system can come to some reliable conclusion using less effort, then there is good reason to think that it would select this route instead of a more effortful one. I am suggesting that if two processes both compute successful attributions with decent probability, then a person not motivated to activate one process over another with regards to accuracy will use the mechanism requiring the least amount of precious working memory resources.

Default processes, or automatic, standard processes seem to follow a pattern: using the least amount of resources but still retaining decent reliability, otherwise another process might intervene[xxviii]. If this is true, and if a system has two processes available, and both reliably mindread, with one requiring less resources than the other, then the system will choose the least effortful process. Because individuals wish to conserve cognitive resources or do not wish to exert them in a given case, I am proposing we can identify the default mechanism of mindreading in this way: if the processes are available and reliable, such that the result of the process is true most of the time, then the relatively less effortful process will be the one deployed most of the time, i.e. default. I believe this line of argument has not been pursued to illuminate the default method of attributing mental states to targets.

Most philosophers hold a hybrid account, one that includes both theory and simulation as viable mechanisms to attribute higher-level mental states to others, but my bi-level theory, in conjunction with the effortlessness approach to identifying primacy, offers a new interpretation for the phenomena. Theory is traditionally thought of as propositional information regarding the mental habits of others, or a folk-psychology; but there is another way of understanding theory—as a connectionist network, which represents information via the strengths among nodes within the network such that some input results in an output based on the regularities of the environment and the strengths of those weights representing those probabilities. While I think we should understand theory in this broader sense to include connectionist networks, I do not think that the existence of the connectionist network can reduce theory to only this neural network. Instead I propose a Dual Type Theory that incorporates both types of theory based on the evidence that our learning of new theories can affect our intuitive, automatic mindreading attributions. Under this account, we have automatic mindreading that is not as intellectualized as traditional theory

yet still capable of being influenced by the new theories we learn because of the interaction between tt1 and tt2 through dual-interventionism and through the learning, or modularization, of new information from tt2 into tt1. Tt1 provides us with a modularity of mindreading, so that humans have an innate capacity to mindread, have an effortless tool to accomplish it, and have an ability to automatize new theories into our intuitive attributions.

From the discussion of the type 1/type 2 distinction, we know that type 1 processes are less effortful, as those types are often referred to as fast, automatic, and "effortless;" while type 2 descriptions include things such as requiring attention, consciousness, etc. as well as features like: being slow, serial, and effortful— but what exactly is "effort?" Perhaps a process can be called "effortful" when it requires attention or cognitive resources as these things are grouped together and may just refer to different aspects of the same entity[xxix]. Type 2 processes use working memory, require attention, are effortful, and conscious; and these different aspects may influence the others, that is it is effortful because it requires attention, and attention uses working memory resources. If this is the case, then type 1 processes require less cognitive resources because they occur without the need for attention, consciousness, or working memory. Type 1 will always run; suggesting that this is always a potential process and could be default since it is does not rely on some input from high-level control systems (Stanovich 2011). It seems to me that a requirement of a default process should be easy accessibility such that the process can be used in every case. As seen with heuristics and cognitive problems like the bat and ball from Kahneman (2011), people utilize type 1 processing in most situations, which is why we see some instances of incorrect intuitive responses when we fail to check our automatic answer[xxx].

Our connectionist network theory is obviously type 1 and traditional theory a type 2 process, but what about simulation? Simulation involves cognitive decoupling so that we can

separate our genuine states from our pretend ones and use imagination to simulate the target's mental state. It also relies on the "like-me" premise to trigger the simulation, and therefore does not automatically engage when there is a stimulus. It relies on information from memory to produce accurate simulations; and it uses resources to inhibit states (as seen in self-regulation). Cognitive decoupling, is a defining characteristic of type 2 processes and relies on working memory, so from the features of simulation, it seems that it can be categorized as a type 2 process, or perhaps a hybrid of type 1 and 2. If so, then mindreading via simulation requires a good amount of resources; therefore, an argument for the primacy of tt1 more likely since it is the least effortful.

While our automatic intuitive responses get it wrong sometimes, for the most part they allow the subject to engage in her environment reliably by responding to cues and statistical probabilities. Both theory and simulation have limitations on accuracy, and therefore have similar reliability; their accuracies depend on other features and for the most part get it right to a similar degree. Tt1 maps the regularities found in the environment, and since it adjusts the weights within the connectionist network to represent the statistical probability of certain stimuli and resulting outputs, it is reliable in constant environments. Tt1 learns bottom-up, but it is also influenced, through supervised learning, by the rule-like theory found in tt2; so the accuracy of the tt1 mindreading also depends on the theory used to shape the tt1 net. It should be reliable if the tt2 is as well since over time it can reflect the knowledge found in tt2. If a person believes a bad theory, she will come to an inaccurate attribution if she directly consults tt2 but also from her intuitive responses that have been affected by her conceptual tt2 knowledge. Simulation is not without inaccuracy either; if the starting states that are fed into the offline module do not reflect those of the target, then the output from the simulation will not be accurate. Another factor of

accuracy for simulation involves the difficulty associated with inhibiting one's genuine states such that they do not contaminate the simulation. If quarantine failure occurs and her states are not separated from the pretend states, then a person will probably attribute incorrect mental states to the target.

Holding that all other aspects are constant—a task that does not differ greatly from our usual experience or environment, etc.—if there are multiple mechanisms available to provide the subject with information about the mental states of others (tt1, tt2, and simulation), with some requiring more cognitive resources (tt2 and simulation), then the mind will default to the one with less strain on limited resources (like attention and memory). Unless motivated, a subject will likely rely on the type 1 theory, but type 2 has the ability to intervene if something triggers it or she is motivated to use the resources; likewise, she can be motivated to use simulation, and she relies on simulation when a theory is not present.

Learning

Because it requires the least amount of resources and automatically triggers, tt1 is the primary mechanism used by people to mindread, that is the one we engage when all processes are possible and reliable. It is modular and provides us with an innate capacity to mindread, but the accuracy of mindreading can be affected by tt2 and simulation. Tt1 can learn from the conceptual knowledge in tt2, such that the theory of tt2 becomes our automatic response. How exactly would this work?

Our innate capacity for mindreading adjusts the weights among the nodes of the connectionist module in response to regularities of environment and successful attributions accordingly. As our networks fine tune these connections through Hebbian learning, that is the increasing of weights between nodes that are activated together and decreasing the weights of

those that are not, learning within tt1 is not limited to this unsupervised kind. Supervised learning is also possible within tt1via the conceptual information of tt2, such that the continuous presentation of the conceptual knowledge can affect the connections and strengths of tt1. Usual learning within type 1 processes are slow, that way the information represents the typical properties of an environment over time (Smith & Decoster, 2000). If a network was altered by just one trial, it could experience an abnormal case, and the resulting change in the network would not be an accurate representation of environmental regularities. This worry is avoided when learning occurs over many trials. On the other hand, type 2 processes can learn quickly to single instances. The two systems can interact through consolidation, which transfers knowledge from the fast process to the slow-learning process through repeated presentations (McClelland et al., 1995); the network still learns over time but via regularities of the type 2 process, not the environment.

According to Baars (1997), consciousness makes propositional content available to the modules by bringing the content to the global workplace. Tt2 is a type 2 process, so it is slow and conscious. If it is conscious, then it has the ability to bring the conceptual content to all the modules via the global workplace. Tt1 is a type 1 process, and being modular, quick, serial, etc. it is one of the modules that conscious content is made available to. As with all modules, if one is susceptible to supervised learning, then this information will affect how it computes. In the case of mindreading, the conceptual tt2 knowledge will act as examples of correct input and output computations such that the connectionist network will alter the weights over time to reflect that knowledge. The learning is possible since tt1 is not content-nonconceptual, but synchronically state-nonconceptual and can be diachronically conceptual. It is not that the adjusting of weights requires attention and effort, for learning in tt1 occurs automatically in response to activation of

nodes, but selecting the correct output through the conceptual knowledge of tt2 does. The process of learning is automatic within the connectionist network, but the desired output that exhibits the theory to be modularized must be provided through attention and consciousness to be accessible to tt1. Over time, the weights will increase among the provided outputs, and less effort will be needed to provide tt1 with the correct output. For example, consider trying to form a new habit or break an old one, such as remembering to turn lights off when exiting a room. At first, one has to remember to turn the light off when she leaves the room, she has to keep that information available in her working memory in order to execute it because she normally does not do it. So if it was not in her working memory she would leave like normal without turning it off. But by attending to that desire, by allocating endogenous attention to it, she is selecting it as a viable behavior. After she executes it and turns the light off, that behavior becomes more likely in the future by adjusting the weights among her corresponding connectionist network. Over time, the amount of attention needed decreases and the action becomes habit, such that she no longer must remember or make a conscious effort to turn the light off, now it has become automatic. Tt1 is not just implementing tt2; while it is reflecting that knowledge and is diachronically conceptual, it is not sentence-like, rule-based information that a classical model proposes. So our tt1 comes to represent the regularities found in both our environment and tt2, such that our conceptual knowledge regarding other minds can influence our automatic attributions, while a simulation-based account cannot explain this.

Developmentally speaking, a modular tt1 provides the capacity to mindread, and more experience and samples of data from the environment will increase the accuracy of this process. Young children often over-generalize based on similar features among stimuli while they are learning which features are associated with one another. The neural nets are still fine-tuning the

weights among the connections to reflect regularities, and often they attribute mental states to others incorrectly because they automatically attribute some mental state they have experience with since they do not have fine distinctions between stimuli. Discussed earlier in the section about type 1 and 2 processes, type 1 process do not differ among individuals based on cognitive ability; however, there seems to be a variation of ability among mindreaders. We can explain this difference by citing the attention and working memory resources necessary to train our tt1 to reflect our tt2 knowledge. If we learn a new theory but cannot train our tt1 accordingly, the our tt1 might still be inaccurate. Working memory and attention do vary with intelligence, so those with higher intelligence will be able to correct and adjust tt1. Other individuals may be executing tt1 networks guided by wrong/incomplete theories.

Simulation

Accuracy of mindreading depends on the connections of tt1 to reflect regularities of the environment and our tt2 knowledge, but it can also be aided by simulation. As previously stated, this is a hybrid account of mindreading; and so both simulation and theory are mechanisms used to mindread, however tt1 is primary. If a child or adult approaches a novel situation and does not have the relevant tt1 knowledge, then accuracy is at risk. Since she is motivated, she will use cognitive resources if she has them. Because simulation requires less resources than tt2, she will engage in simulation if she does not possess a reliable tt1 network. If simulation is not possible for whatever reason, she can accept the automatic tt1 response, which most likely has experience with something similar to the novel case, or she could access tt2 if she has cognitive resources but unable to simulate because of other factors. If we are lacking a folk-psychology regarding this scenario, then we must engage in mental simulation for an accurate attribution. Simulation can also add to our theory and thus increase accuracy of mindreading in general. Once the person

engages in mindreading and concludes an attributable mental state, that mental state is tagged and projected to the other person—but it is also tagged and generalized to potentially be added to the tt2 knowledge regarding the minds of other people. The mechanism responsible for this task is the same self-reflection proposed by Goldman to attribute and tag states as metarepresentational (2006). As children go through experiences and learn about self, they can tag that knowledge as regarding other minds as well; so the foundation of our tt2 is based on our own experiences and generalized out. That new knowledge (tt2) can now be made available to the modules, as described above, and train tt1 over time. The generalization of our own psychology and experiences explains why our theories are laden with egocentric tendencies and why we exhibit egocentric bias, the demonstration of which has led some to believe that simulation is responsible. But as I have shown, egocentric bias and other mindreading phenomena are compatible with a Dual Type Theory Account, an account that incorporates two types of theory and argues for the primacy of the tt1 mechanism.

Conclusion

Attributing high-level mental states to others is a complex phenomenon that incorporates both theory and simulation; however, one process is likely used more than the other. Authors have argued for theory-based and simulation-based accounts, but my Dual Type Theory Account provides a new approach to understanding aspects of the mindreading phenomenon while providing reasons to accept theory as the default, or primary, mechanism. By dividing theory into a traditional database of conceptual knowledge (tt2) and connectionist networks (tt1), we can explain a multitude of features and weaken arguments previously used as evidence for a simulation-based account. For example, self-reflection and egocentric bias, both of which are prevalent in mindreading, can now be explained on a two-type theory-based model, whereas they

were not obviously implied by previous accounts of theory. At the same time, we avoid the over-intellectualized worry about a traditional folk-psychology. With the tt1, we have a modular mindreading capacity that is still susceptible to supervised learning, thus we can explain why our attributions are innate but can be cognitively penetrated.

Furthermore, the division of theory into different types illuminates features of each mechanism, such as the effort required of each. As I have shown, tt1 is the least effortful mindreading process since it is fast, automatic, and does not rely on working memory resources. On the other hand, the remaining mechanisms require more resources and are not automatically engaged. There is reason to believe that individuals select the least effortful process most of the time; and with regard to mindreading, that process would be tt1. Therefore, there is reason to believe that tt1 is our default mindreading mechanism on a Dual Theory Type Account.

In addition to the current conversation regarding primacy in mindreading, the Dual Type Theory Account has many important implications that I will briefly mention. We now know that our automatic responses when mindreading rely on our experiences with the environment but can also be guided by tt2 knowledge. When we are motivated, our tt2 can intervene with our tt1 processes, but our tt2 knowledge can also adjust our connectionist networks using attention and working memory. Therefore, we may draw some practical, albeit speculative, applications from this new account of mindreading.

There could be a potential opportunity to help individuals improve their mindreading abilities by teaching correct theories about others' minds and encouraging supervised learning, so that the new knowledge becomes modularized and automatic. We can motivate individuals to access the tt2 knowledge during cases of mindreading and to train their tt1 networks. New theories are learned from a single instance within tt2, but our tt1 requires more trials. However, if

we motivate or remind individuals to access the tt2 information during the mindreading process, then default-interventionism may provide the subject with a more accurate attribution. Additionally, we could instruct individuals to reflect on past attributions and compare them to the new tt2 knowledge. Over time, the tt1 networks would adjust. Motivation could be present by addressing the past theories as obsolete and incorrect (since the best motivator to engage type 2 processes is accuracy).

Similarly, we have a picture of the causes and solution to implicit bias. Implicit bias occurs when people unconsciously hold biases towards members an outsider group. These types of judgments suggest that they are not engaging in simulation, since they do not attribute things that they would attribute to themselves if they were in that situation. Additionally, they would not engage in simulation because they feel so removed from the target that it would be a waste of resources. So how do they make those judgements? They are relying on their tt1 networks—automatic responses that have been shaped by their environment (cultural behavior) and tt2 knowledge regarding some group of people. The more that theory is confirmed, say by a demagogue who preaches racial prejudice, the stronger that connection becomes in the tt1 of individuals. Implicit bias is cultural, thus supporting the hypothesis that it is a result of some tacit theory; however, we potentially have a solution to correct this wrong attribution. We can teach correct theory (tt2) and provide instructions to access that information over many trials of outsider group mindreading. But what I think would be more instrumental is to induce simulation, in which biased individuals simulate the situation of an outgroup person. The aversion to simulate with people "unlike us" stops individuals from engaging in this mechanism. However, we can provide vignettes and instruct subjects to simulate the situation and mental states of the target, whose status, race, gender, etc. is unknown. After the subject engages in

simulation, she would be informed that the target was a member of an outsider group. The illusion of "unlike me" would weaken and the simulated output could be generalized and added to tt2, which then can adjust tt1. This process must, again, be motivated, but has the potential to slowly correct our biases.

Although these practical implications are only speculative, they demonstrate that the Dual Type Theory Account provides opportunities to improve and correct mindreading. By understanding the different mechanisms and how they interact, we also get a sense of how to increase our accuracy. These solutions and therapies are not readily available on previous accounts of mindreading and rely on the division of theory into two types, that is an automatic innate capacity with the ability to alter intuitive responses. While we cannot really know what someone else is thinking, we can become better equipped at attributing mental states to others and interact with one another more efficiently and productively.

References

Baars, B. J. (1997). *In the theater of consciousness: The workspace of the mind.* New York, NY: Oxford University Press.

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological validity to dual processes. *Behavioral and Brain Sciences*, *30*, 241–297.

Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In J. R. S. Wyer, & T. K. Srull (Eds.), *Handbook of social cognition* (Vol. 2, pp. 1-40). Hillsdale, NJ: Lawrence Erlbaum Associates.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. Cambridge, MA: MIT Press.

Chaiken, S., & Trope, Y. (Eds.). (1999). *Dual-process theories in social psychology*. New York, NY: Guilford.

Chalmers, D. J. (1990). Why fodor and Pylyshyn were wrong: The simplest refutation. In *Proceedings of the twelfth annual meeting of the cognitive science society,* 340-347. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Churchland, P. (1989). Folk psychology and the explanation of human behavior. In P. Churchland (Ed.), *A neurocomputational perspective* (pp.111-128). Cambridge, MA: MIT Press.

Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.), *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control* (pp. 62-101). Cambridge, UK: Cambridge University Press.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*(4), 450-466.

Epley, N., Keysar, B., Van Boven, L. & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality & Social Psychology, 87*, 327-39.

Evans, G., 1982. *The Varieties of Reference*. Oxford, UK: Oxford University Press.

Evans, J. St. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgement*. Hove, UK: Psychology Press.

Evans, J. St. B. T. (2010a). Intuition and reasoning: A dual-process perspective. *Psychological Inquiry*, *21*, 313–326.

Evans, J. St. B. T. (2010b). *Thinking twice: Two minds in one brain*. Oxford, UK: Oxford University Press.

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223-241.

Fodor, J. (1987). *Psychosemantics*. Cambridge, MA: MIT Press.

Goldberg, L. S., & Grandey, A. A. (2007). Display rules versus display autonomy: Emotion regulation, emotional exhaustion, and task performance in a call center simulation. *Journal of Occupational Health Psychology, 12*(3), 301-318.

Goldman, A. I., & Shanton, K. (*forthcoming*). The case for simulation theory. In J. A. Leslie & T. German (Eds.), *Handbook of 'theory of mind'*. New York, NY: Taylor and Francis Group.

Goldman, A. I. (2006). High-level simulational mindreading. In A. I. Goldman (Ed.), *Simulating minds: The philosophy, psychology, and neuroscience of mindreading* (pp. 147-191). New York, NY: Oxford University Press.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109*(1), 75-90.

Gordon, R. (1995). Simulation without introspection or inference from me to you. In T. Stone and M. Davies (Eds.), *Mental simulation*. Oxford, UK: Blackwell.

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

Grillon, C., Quispe-Escudero, D., Mathur, A., & and Ernst, M. (2015). Mental fatigue impairs emotion regulation. *Emotion, 15*(3), 383-389.

Heck, R. G. (2000). Nonconceptual content and the space of reasons. *Philosophical Review*, *109*, 483–523.

Heck, R. G. (2007). Are there different kinds of content? In J. Cohen, & B. McLaughlin (Eds.), *Contemporary debates in the philosophy of mind* (pp. 117-138). Oxford, UK: Blackwell.

Hilbig, B. E., & Pohl, R. F. (2008). Recognition users of the recognition heuristic. *Experimental Psychology, 55*(6), 394-401.

Hofmann, W., & Timothy, D. W. (2010). Consciousness, introspection, and the adaptive unconscious. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications*, (pp. 197-215). New York, NY: Guilford Press.

Kahneman, D. (2011).*Thinking, fast and slow*. New York, NY: Farrar, Straus, and Giroux.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgement. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge, MA: Cambridge University Press.

Johns, M., Inzlicht, M., & Shmader,, T. (2008). Stereotype threat and executive resource depletion: Examining the influence of emotion regulation. *Journal of Experimental Psychology: General, 137*(4), 691-705.

Keysar, B., Lin, S., & Barr, D.J. (2003). Limits on theory of mind use in adults. *Cognition, 89*, 25-41.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102, 419-457.

Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex. *Journal of Cognitive Neuroscience, 17*(8), 1306-1315.

Nichols, S. (Forthcoming). Mindreading and the Philosophy of Mind. In J. Prinz (Ed.), *The oxford handbook on philosophy of psychology*. New York, NY: Oxford University Press.

Nisbett, R., & Ross, L. (1980). *Human inference*. Englewood Cliffs, NJ: PrenticeHall.

Tager-Flusberg, H., & Sullivan, K. (2000). A componential view of theory of mind: Evidence from williams syndrome. *Cognition, 76*, 59-89.

Thompson, V. A., Turner, J. P., & Pennycock, G. (2011). Intuition, reason and metacognition. *Cognitive Psychology*, *63*, 107–140.

Sellars, W. (1963). Empiricism and the philosophy of mind. In W. Sellars (Ed.), *Science, perception and reality* (pp. 127-196). London, UK: Routledge and Kegan Paul.

Scholl, B., & Leslie, A. (1999). Modularity, development, and 'theory of mind'. *Mind & Language*, *14*, 131-153.

Schweizer, S., Grahn, J., Hampshire, A., Mobbs, D., & Dalgleish, T. (2013). Training the emotional brain: Improving affective control through emotional working memory training. *Journal of Neuroscience, 33*(12), 5301–5311.

Smith, E. R., & DeCoster, J. (2000). Dual process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review, 4*(2), 108-131.

Smolensky, P. (1988a). On the proper treatment of connectionism. *The Behavioral and Brain Sciences, 11*(1)*,* 1-23.

Smolensky, P. (1988b). Connectionist mental states: A reply to fodor and pylyshyn. *Southern Journal of Philosophy, 26,* 137-161.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.

Stanovich, K. E. (2004). *The robot's rebellion: Finding meaning the age of darwin*. Chicago, IL: University of Chicago Press.

Stanovich, K. E. (2011). *Rationality and the reflective mind*. New York, NY: Oxford University Press.

Stich, S., & Nichols, S. (1992). Folk psychology: Simulation or tacit theory? *Mind & Language, 7*(1), 35-71.

Stich, S., & Nichols, S. (2003). Folk psychology. In S. Stich & T.A. Warfield (Eds.), *The blackwell guide to philosophy of mind* (pp. 235-255). Oxford, UK: Blackwell.

Xie, W., & Zhang, W. (2016). The influence of emotion on face processing. *Cognition and Emotion, 30*(2), 245-257.

Footnotes

[i] Although these seem like very different activities, both are included in the mindreading phenomenon.

[ii] The Dual Type Theory is distinguishing between two processes within theory: type 1 and type 2. This distinction is the same as the classic partition of system 1/system 2: type 1 possessing the same cluster of features as system 1 and type 2 involving the qualities usually attributed to system 2.

[iii] By complex emotion, I mean emotions that are not obvious from only a subject's perception of the target's facial expression. "Complex" meaning dependent on some circumstance or social quality that is not directly accessible through perceptual appearance alone. In this category are things such as anxiousness or suspicion. I also mean emotions about certain things, or attributing reasons for emotions, like being sad about X, as opposed to identifying a person as only sad.

[iv] Appealing to the effort required by each mechanism has not been used as evidence for one over the other. However, the proposed division of theory invites such an argument, as the different types have varying demands on cognitive resources.

[v] For the most part. Some beliefs are not transparent such as ones we have compartmentalized, blocked out, or possess unconsciously like implicit beliefs.

[vi] This distinction between ST and TT as knowledge-poor versus knowledge-rich is rejected by Goldman (2006). He stated that "there is no reason that creation of pretend states should not rely on information stored in memory. This does not prevent a cognitive operation from being a simulation" (p. 150). What makes it simulative is the top-down production of a state with the purpose of replicating a normally produced state.

[vii] False-belief task is also evident in adults, but more so in children.

[viii] While simulation does not rely on a vast database of conceptual knowledge, it does require cognitive resources more so than my proposed tt1, to be discussed further.

[ix] The nodes and connections among them are analogous to neurons and the synapses respectively. The weights affect the synapses, such that activated neurons could activate or inhibit another depending on the activation value of each and weights of the connections to that neuron. This process continues from input nodes to hidden nodes to output nodes, which provide the network with an outcome or decision. It is not a simple association principle, but differential equations that guide the system. See Smolensky, 1988a.

[x] The outputs are selected as inputs for some other mechanism like the projecting of a mental state and tagging it as a metarepresentational belief.

[xi] Not only are connectionist networks responsive to environmental regularities, but to our tt2 conceptual information, which can alter the weights of a network over time, so they reflect the information represented within tt2 but accessed automatically and with the least amount of effort.

[xii] Known as the "position effect," individuals will evaluate the two identical items differently, such that the item on the right was preferred to the one on the left. See Nisbett & Ross, 1980, p. 207.

[xiii] Previously referred to as "system 1/system 2" in the literature, "type" is more appropriate by remaining agnostic about how many systems there are, as opposed to suggesting that there are only two. For instance, there are many systems (visual system, auditory system) in "system 2."

[xiv] The types often possess these characteristics but do not necessarily need to in order to be classified as one or the other. On the other hand, defining features are ones that a process must have to be categorized as type 1 or type 2.

[xv] A bat and a ball cost $1.10. The bat costs one dollar more than the ball. How much does the ball cost?" Most people intuitive answer ¢10 because we quickly subtract $1 from the total, however the correct answer is ¢5, but requires type 2 intervention. See Kahneman, D. (2011).

[xvi] The output is not affecting the subject's states because it has been taken offline, but it can be selected as an input for another task.

[xvii] The database of traditional TT relies on vast conceptual knowledge that is not innately given and appears too intellectualized.

[xviii] These different tasks are not as easily accomplished as ST suggests. They are not as obviously accessible as the mechanism used to extract information from TT theory.

[xix] However, self-reflection is expected on my proposed view.

[xx] Stating the downfall of the process as difficult may be evidence against it being the primary mechanism, rather than for it. Default mechanisms are the ones that occur most often, and there is reason to believe that a process of this type would be effortless.

[xxi] Emotion regulation is relevant to discussions of high-level, simulative mindreading as we must often inhibit our own emotion states when simulating another's intense sadness for a deceased parent, for example. Therefore, by showing that emotion regulation requires cognitive resources, inhibition must also require resources.

[xxii] We will see, in the hybrid section, that the role of self-reflection to tag states is also compatible with theory mindreading. Therefore, its appearance during mindreading tasks is not obvious evidence for ST.

[xxiii] That is, cases that are not motivated to activate an additional process to intervene the intuitive response.

[xxiv] Phylogenetically speaking, Goldman proposes that ST is more in line with what evolution likely created. Since simulation consists of using one's own modules as the model of the target by running them offline to conclude an output to be projected onto the person in question, some might consider simulation to be phylogenetically more plausible, and so primary, at least Goldman does (2006). He states that creating a new module is more work than using the ones we have in a different way, and so it is more likely that we just use our own thought patterns and project them onto others instead of developing two separate theories (one for our own mind and one for the minds of others). However, it is not clear how this is an argument for simulation on the basis that it would be easier for evolution (or less work), when simulation requires cognitive resources that could be retained if using tt1.

[xxv] Another point by Goldman concerns the ontological aspect of both theories. Acquisition is not as obvious for traditional TT as that of ST, and Goldman takes this ontological problem to be stronger evidence for ST as default. Since beliefs concerning myself do not constitute a folk-psychology, but can guide simulation, Goldman states that acquisition makes more sense for simulation. However, even small children are capable of creating theories that reflect, for example grammar or science, and they acquire the relevant knowledge even quicker than knowledge of folk-psychology (Stich & Nichols, 1992). Moreover, I propose that our tt1 encodes the regularities we encounter with other minds and so develops in that way, fine tuning with experience, but also is susceptible to our conceptual knowledge of folk-psychology. And our conceptual knowledge can include generalizations from our own experiences and our simulations of other people's mental states. So acquisition with a Dual Type Theory is explainable and not as difficult as Goldman suggests.

[xxvi] See note xv.

[xxvii] This type of reflection is not necessarily conscious.

[xxviii] Such as seen in default interventionism, when type 2 processes can intervene the intuitive, type 1 process.

[xxix] As stated earlier, this is open for discussion but not pursued here.

[xxx] This is not suggesting that there is some increased fallibility associated with all Type 1 processes that is absent from Type 2 processes. On the contrary, there are times when our Type 1 processes can lead to accurate answers while Type 2 might conclude biased responses (see Evans, (2007); Stanovich, (2011); as cited in Evans & Stanovich, (2013)).

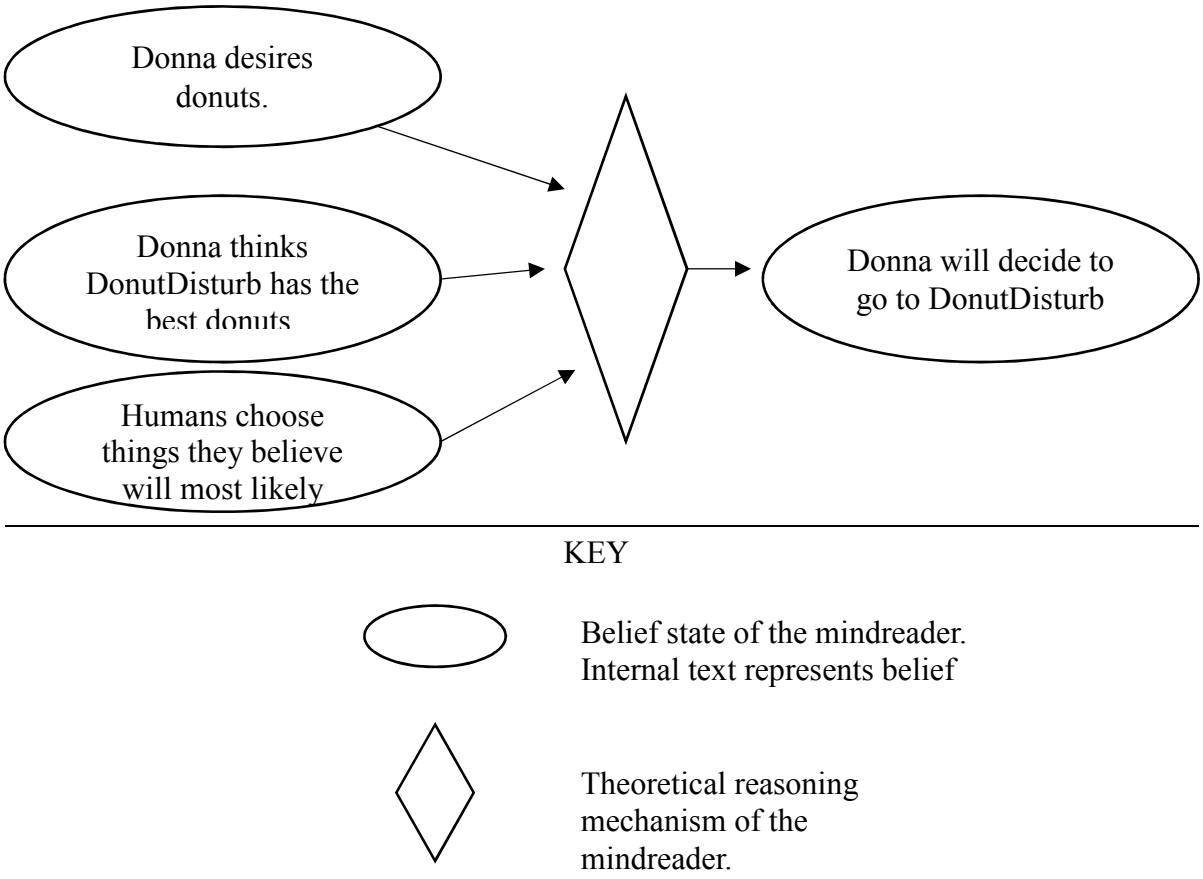

Figure 1. TT mindreading process. Based off the reconstruction by Goldman & Jordan (2013).

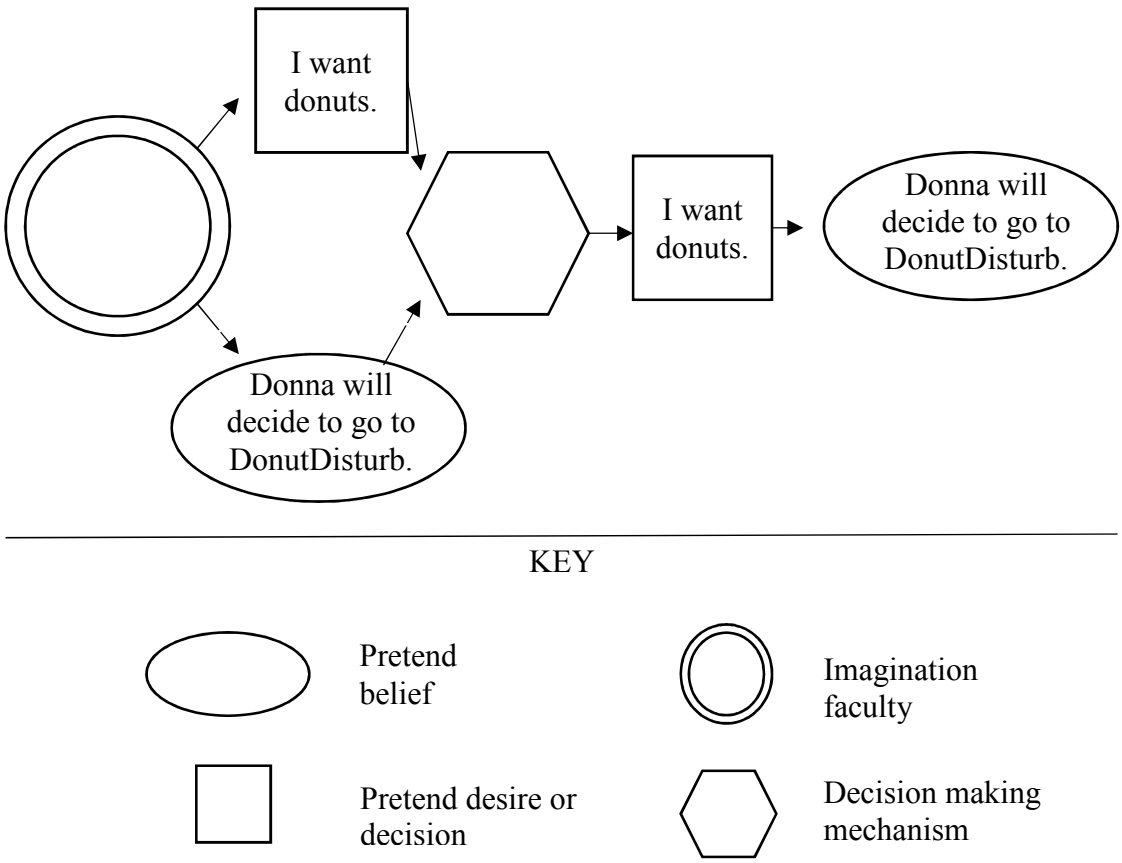

Figure 2. ST process of mindreading. Based on the model provided by Goldman & Jordan (2013).