Fisher, Justin. (2006) Does Simulation Theory Really Involve Simulation? *Philosophical Psychology* 19: 1–16.

Does Simulation Theory Really Involve Simulation?

Justin C. Fisher – Southern Methodist University

Abstract.

This paper contributes to an ongoing debate regarding the cognitive processes involved when one person predicts a target person's behavior and/or attributes a mental state to that target person. According to Simulation Theory, a person typically performs these tasks by employing some part of her brain as a simulation of what is going on in a corresponding part of the brain of the target person. I propose a general intuitive analysis of what simulation means. Simulation is a particular way of using one process to acquire knowledge about another process. What distinguishes simulation from other ways of acquiring knowledge is that simulation requires, for its non-accidental success, that the simulating process reflect significant aspects of the simulated process. This conceptual work is of independent philosophical interest, but it also enables me to argue for two conclusions that are of great significance to the debate about mental Simulation Theory. First, I argue that, in order to stake a non-trivial claim, Simulation Theory must hold that mental simulation involves what I call concretely similar processes. Second, I argue for the surprising conclusion that a significant class of cases that Simulation Theorists have claimed as intuitive cases of simulation do not actually involve simulation, after all. I close by sketching an alternative account that might handle these problematic cases.

Keywords: Simulation Theory, Folk Psychology, Mental State Attribution, Mindreading

1. Introduction.

This paper contributes to an ongoing debate regarding the cognitive processes involved when one person predicts a target person's behavior and/or attributes a mental state to that target person. According to one popular position in this debate, Simulation Theory, a person typically performs these tasks by employing some part of her brain as a *simulation* of what is going on in a corresponding part of the brain of the target person.[1] One conclusion of this paper is that, surprisingly, a large number of cases that Simulation Theorists have counted as intuitive cases of simulation turn out, under closer inspection, not actually to involve simulation, after all.

[1] For a good overview of this debate and of Simulation Theory, see the introduction to Davies and Stone (1995).

To reach this surprising conclusion, I begin by proposing a general intuitive analysis of what 'simulation' means. Simulation is a particular way of using one process to acquire knowledge about another process. What distinguishes simulation from other ways of acquiring knowledge is that simulation requires, for its non-accidental success, that the simulating process *reflect* significant aspects of the simulated process. My analysis of simulation is intuitively satisfying, and that it is roughly what Simulation Theorists, at least in some instances, have thought simulation to be. This conceptual work is of independent philosophical interest, but it also enables me to argue for two conclusions that are of great significance to the debate about mental Simulation Theory. First, I argue that, in order to stake a non-trivial claim, Simulation Theory must hold that mental simulation involves what I call *concretely similar* processes. Second, I argue for the surprising conclusion mentioned above – that a significant class of cases that Simulation Theorists have claimed as intuitive cases of simulation do not actually involve *simulation*, after all. I will close by sketching an alternate account that might handle these problematic cases.

2. Processes.

My first goal will be to unpack and defend my general analysis of simulation. I draw upon Goldman's (2003, in press) idea of setting up the analysandum in terms of one process simulating another. My proposal is as follows:

A *process* P_1 is a ***simulation*** of some other (perhaps *hypothetical*) process P_2 iff

(1) a *purpose* of P_1 is to facilitate *knowing* about P_2; and

(2) this particular purpose can be fulfilled in a *non-accidental* way only if P_1 *reflects significant aspects of* P_2.

A simulation is a *process*, and that what a simulation simulates is a *process* as well. A ***process*** is a sequence of successive states of a system which is such that its state at any given time *largely determines* its state at immediately subsequent times.[2] Examples include the successive states of the U.S. economy, of an airplane in flight, or of a thinking mind.

As an idealization and formalization, let $N_S(x)$ be the function which yields the state of system S that is normally subsequent to (or 'next' after) state *x*. Additionally, for any system that will be a candidate for being simulated, we should define a "similarity" relation '≈' that may obtain between two states that are sufficiently *similar* by some theoretically appropriate measure of similarity.[3] If system S is a good candidate for being simulated, the relation '≈'should be *mostly transitive* (i.e., for most states *x, y* and *z* of S, $(x \approx y \ \& \ y \approx z) \supset x \approx z$), and the function $N_S(x)$ should preserve '≈' in most cases (i.e., for most states *x* and *y*, $x \approx y \supset N_S(x) \approx N_S(y)$).[4]

3. Epistemic Purposes.

Simulations are epistemic devices. The purpose of doing a simulation is to produce knowledge about the process that is being simulated. A process that *just happens to* mirror another process – however well – is not a simulation. An imitation done *solely* for the sake of imitation is not a simulation either. Many such processes may be *potential simulations* – they

[2] Intuitively, the U.S. economy is driven by both *economic factors* (e.g., interest rates and consumer confidence) and *non-economic factors* (e.g., droughts, political coups, random fluctuations). The success of economics as a science depends upon using *economic factors* to predict economic changes, and depends upon the fact that *non-economic factors* normally play a limited or predictable role in causing economic changes. This is what I have in mind by saying each state ***largely determines*** the next – not that the next state actually is a function *solely* of the preceding state, but just that it is close enough that good explanations might be given for *why* a system normally enters the subsequent states *in terms of* the system's previous states.

[3] Which measure is theoretically appropriate may depend upon our purposes in attempting to understand a system as a case of simulation.

[4] In addition, we would probably want '≈' to be *reflexive* (for any state *x*, $x \approx x$), *symmetric* (for any states *x* and *y*, $x \approx y \supset y \approx x$), though these further constraints will play no role in the argument below.

may be such that they might easily be employed *as* simulations. But, unless a process is employed for an epistemic purpose it is not employed as a simulation.[5]

I intend both '*purpose*' and '*epistemic*' quite broadly. For example, I want to allow that when young animals play with one another, they may be engaging in *simulations* of actual combat (or of actual predator-prey dynamics) for the purposes of learning how best to react in such situations. Young animals do not have *intentions* to derive such benefits from their play, but still there is a clear sense in which their play *does have* this purpose – specifically, evolution selected for playing *because* playing is enough like combat to prepare animals for it.[6] Animals almost certainly do not acquire *propositional knowledge* from their play, but what they acquire is surely some form of *knowledge*. (It may be *know-how* rather than *know-that*.) Nor is play absolutely reliable at giving young animals the sort of knowledge they need – instead it just does well enough (from an evolutionary perspective) for it to be worth doing.

This raises an important point. Simulations are generally not perfect. Many simulations may not even fulfill their purposes at all – they may fail to provide any sort of accurate knowledge of the process being simulated.[7] Imperfection and failure are consistent with the idea of a purpose. One purpose of a mouse's auditory system is to let it know of approaching predators, but many mice do get eaten, and some mice are deaf. A good theory of mouse auditory systems – just like a good theory of simulation – must leave room for imperfection and outright failure. But on the other hand, to best understand a mouse's auditory system, we must

[5] Common intuition may count as simulations such things as 'flight simulators' or Virtual Reality systems that are purely for *entertainment purposes*. These cases do not involve *epistemic* purposes, but they do admit *evaluations* as to how successful the simulation is, and why. Probably, this *evaluability* is actually what is most generally required for a simulation, but in most central cases this evaluability derives from the *epistemic* uses a simulation is put to. I will concentrate on those central cases here.

[6] The sense of natural purpose I have in mind here is similar to Millikan's (1984) 'proper functions,' and indeed, the depiction of simulation I offer here shares much of the spirit of Millikan's work. However, the theory of simulation proposed here should be consistent with other theories of 'purpose' as well.

[7] This point is stressed by Goldman (in press).

attend to the cases in which this system works as it is supposed to. Similarly, to best understand simulations, we should attend to those cases when all goes well, those cases in which a simulation non-accidentally succeeds at its purpose. Let us move on to consider such cases now.

4. System Homomorphisms.

What distinguishes *simulation* from other methods of generating knowledge is the way simulations are supposed to work. A simulation is supposed to work by providing an epistemically-available process that *reflects* the relevant aspects of some process that is not so epistemically-available. We simulate because the simulating process is easier or safer to access than is the target process that we really want to know about. But, in order for this trick to work, it needs to be the case that the simulation we use is *relevantly like* the target process. Ideally, the simulation will accurately reflect the important aspects of the target so that observations of the simulation may yield accurate conclusions about the target.

As a starting point, let's consider Goldman's (1995) suggestion that a simulation must involve an isomorphism. An isomorphism between set S and set T is a mapping φ from S to T, such that (1) φ is 'one-to-one', (2) φ is 'onto', and (3) φ preserves each relevant function f.[8] I think this is much too strong. Clearly, simulation does not require that the mapping be 'one-to-one'. For example, we may easily use a 24-hour clock to simulate a 12-hour clock, even though that involves a two-to-one mapping. We also shouldn't require an 'onto' mapping. We want to allow that a child might passably simulate some simple adult emotions, even if the child entirely lacks the capacity to enter into anything like the adult's more sophisticated emotional states. To allow for this possibility, we will need a weaker notion than *isomorphism*.

[8] More formally: (1) φ associates a different element of T to each element of S; (2) each element of T is associated by φ to some element of S; and (3) for each relevant n-ary function f that operates on elements of S, there is a corresponding n-ary function f^* that operates on elements of T such that for any $s_1, s_2, \ldots, s_n$ in S we get $\varphi(f(s_1, s_2, \ldots, s_n)) = f^*(\varphi(s_1), \varphi(s_2), \ldots, \varphi(s_n))$.

I suggest that the sort of mathematical relation we'll need is a *loose and partial homomorphism*[9] from the simulating system to the simulated system. Suppose system S is being used to simulate a process of system T. First of all, we will need a mapping φ from possible states of system S to possible states of system T. This mapping needn't be defined on *all* states of S, just those that may play into plausibly successful cases of simulation. Let us call the set of states that φ is defined upon the *relevant* states of S – they are the states of S that are relevant to our understanding S as a potential simulative device. For the reasons stated above, we needn't demand that φ be *one-to-one* nor that it be *onto*. For convenience, let us say that state *s* and state *t* ***φ-correspond*** just in case $\varphi(s) = t$.

Now, let us take N_S and N_T to be next-functions (as defined in Section 1) for systems S and T respectively, and recall the similarity relation '≈' that was also defined in Section 1. There is a ***system homomorphism*** from system S to system T just in case there exists a mapping φ as described above, such that for almost[10] any relevant state *s* of S we get

$$\varphi(N_S(s)) \approx N_T(\varphi(s))$$

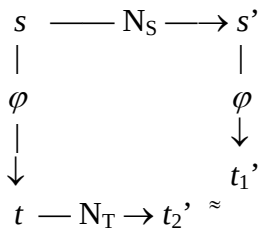
Intuitively, this says that for almost any relevant state *s* of the *simulating* system, you would get very *similar* predictions about the subsequent state in the *simulated* system in two ways: (1) by looking at what in the simulated system φ-corresponds to the state that would come $next_S$ after *s* in the *simulating* system, or (2) by looking at what would come $next_T$ in the simulated system after the state in that system that φ-corresponds to *s*. This is illustrated in Figure 1.

---

[9] A homomorphism from set S to set T is a mapping φ from S to T that preserves each relevant function *f*. By 'partial' I mean that φ needn't be defined on *all* elements of S. By 'loose' I mean that φ needn't *strictly* preserve the function $N_S(x)$, but rather that it need only preserve $N_S(x)$ up to the similarity relation '≈'. This is made clear below.

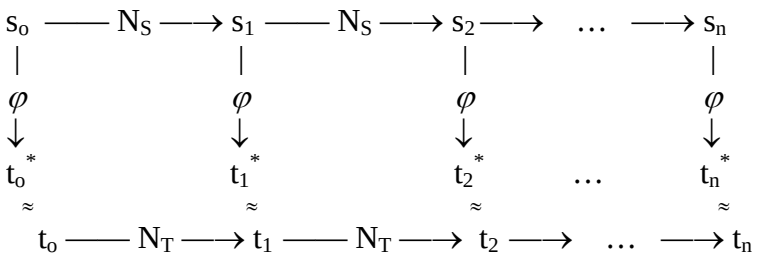
[10] Saying 'almost' allows the possibility that even clear cases of simulation might involve a few consistent errors.

Figure 1. Beginning with state *s*, you get similar predictions (1) by applying N_S and then φ, or (2) by applying φ and then N_T.



So, one condition upon the non-accidental success of a simulation is that there exist a *system homomorphism* from the simulating system to the simulated system. A second condition is that each state in the simulating process *actually does* φ-correspond to a state that is similar to the matching state in the simulated process. (I.e., for each state s_i of the simulating process and each state t_i of the simulated process, it must be that $\varphi(s_i) \approx t_i$.)[11] This is illustrated in Figure 2.

Figure 2. In a non-accidentally successful case of simulation, each state s_i of the simulating process will map to a state t_i^* that is similar to the matching state t_i of the simulated process.



[11] Given that the simulation starts in the right state (i.e., that $\varphi(s_o) \approx t_o$), the definition of *system homomorphism*, and the conditions upon '≈' (given in Section 1), it follows that it is probable that, over a short run, the subsequent states of the simulating process will φ-correspond to states that are similar to subsequent states of the simulated process. Inductive proof for this claim: Base case: We are given that $\varphi(s_o) \approx t_o$, so it is probable that $\varphi(s_o) \approx t_o$. Inductive case: Suppose it is probable that $\varphi(s_i) \approx t_i$. By N_T's preservation of '≈' it is probable, then, that $N_T(\varphi(s_i)) \approx N_T(t_i)$, or equivalently that $N_T(\varphi(s_i)) \approx t_{i+1}$. By the system homomorphism, it is probable that $\varphi(N_S(s_i)) \approx N_T(\varphi(s_i))$, or equivalently that $\varphi(s_{i+1}) \approx N_T(\varphi(s_i))$. By the probable transitivity of '≈', then, it is probable that $\varphi(s_{i+1}) \approx t_{i+1}$. However, the expected probability of this should not exceed the probability that $\varphi(s_i) \approx t_i$, as the inference relied upon other merely probable claims (homomorphism, N_T's preservation of '≈', and the transitivity of '≈'). Hence, over a long run the two processes could not be expected to remain in near φ-correspondence, but (assuming the involved probabilities are high), over a sufficiently short run they can.

Or, to summarize this illustration colloquially, a non-accidentally successful simulation must start off close to correct, and as it runs, it shouldn't get too far off.[12]

5. Concrete Similarity vs. Abstract Similarity.

Now that I have sketched my general analysis of simulation, let me now make precise an important distinction between the two ways in which a simulating system might reflect the causal structure of the system it simulates. Consider the following two examples:

(1) Aerodynamics in the wind tunnel is enough like aerodynamics in open air to allow us to make accurate predictions of how actual planes will fly on the basis of what happens to model planes in the wind tunnel.

(2) A sophisticated computer simulation of air-flow reflects quite well the patterns of actual air-flow – well enough to enable accurate predictions of how planes will fly.

There is a clear intuitive distinction between these two cases. The wind tunnel works because there is a quite ***concrete*** similarity between the aerodynamic processes that affect the model plane and the aerodynamic processes that affect the real airplane. These processes are similar both in the *type of system* that is involved, and in the *fine-grained details* of how that system advances from state to state. The processes at play in the computer simulation, however, are of a quite different sort. The success of the computer simulation is attributable to a shared ***abstract***

[12] One might wonder why I bothered with all the talk of *system homomorphisms* if this is the picture I end up with. Why not hold that this mapping from stages of one process to stages of the other process is *all* it takes to have a case of successful simulation? The answer is that we needed, somehow, to determine what mapping function φ to use. Pictures like Diagram 2 are easy to construct if we aren't picky about the function φ. For example, we could create a mapping from the successive stops on the first cross-continental railroad to the successive holders of the U.S. Presidency, and this would yield Diagram 2. Yet, clearly, this mapping between the two processes *does not mean* that one could be a simulation of the other. The *system homomorphism* criterion guarantees that we can't choose *any old* mapping as φ. More specifically, it guarantees that a non-accidentally successful simulation must involve a mapping that captures the similarity between the causal structures of the involved systems. There is a deep and interesting parallel between this problem and Putnam's (1977) observation that for any model, there will be indefinitely many potential isomorphisms between that model and (suitably gerrymandered portions of) the world. For a good discussion of Putnam's puzzle see Lewis (1984). What I've proposed parallels one solution to this puzzle that Lewis discusses: namely, to constrain the choice of mapping in such a way that an appropriate mapping must capture *real causal relations* rather than just any old gerrymandered relation in the world.

pattern present in the functionality of two quite different systems (computer and flying-plane), systems which reflect one another only in coarse-grained details.[13]

6. Mental-Mental Simulation

Thus far, everything I have said applies to simulations in general. For the remainder of the paper we will be concerned just with simulations that are employed for mentalizing. These are simulations that are employed by (and within) a mind/brain for the purpose of knowing about a mind/brain. Following Goldman (2003, in press), I will call these ***mental-mental simulations***. Mental-mental simulation may be used to yield representations of the mental states of the target and/or predictions of how the target might behave.[14]

Two leading contenders in the debate regarding how people mentalize are Simulation Theory and Theory-Theory. Simulation Theory holds that mentalizing typically involves mental-mental simulation. Theory-Theory holds that typical mentalizing tasks involve primarily the employment of some sort of tacit theory of mind.

Many versions of Theory-Theory entail *some form* of mental-mental simulation. To see this, consider how the Theory-Theorist would account for a simple case. Suppose that Joe knows that Mary has just poured herself a cup of coffee; that Mary likes milk in her coffee; that the refrigerator contains both a half-full container of milk with yesterday's date on it and a fresh

---

[13] Goldman (1995) proposes a distinction between *process-driven* and *theory-driven* simulations which cuts along almost the same lines as the intuitive distinction proposed above. Goldman writes, "If a person seeks to simulate the weather or the economy, in the sense of mentally constructing or anticipating an actual (or genuinely feasible) sequence of its states, she is very unlikely to be accurate unless she has a good theory of the system. A successful simulation of this kind must be *theory-driven* let us say." (Goldman, 1995, pg. 85). This captures what's going on in many actual computer simulations which have been programmed in accordance with theories of the simulated phenomena. But imagine a computer program that was generated at random; or better yet, imagine a program that was selected by some evolutionary process for its capability to make good aerodynamic predictions. A computer program generated in this way may certainly be put to epistemic use *as* a simulation. However, such a program does not embody an aerodynamic *theory*, or if it does, this theory must be an incredibly *tacit* one. Hence, the presence or absence of a *theory* does not quite capture the clear intuitive distinction noted above between simulations like the wind tunnel and simulations like the computer model.

[14] In cases of analysis-by-synthesis, mental-mental simulations might also be used to generate knowledge of the initial conditions that prompted certain of a target's mental processes.

unopened container of milk; and that Mary is extremely wary of food-poisoning. A Theory-Theorist might hold that Joe would predict Mary's behavior by way of an inferential process involving the following highlights: Mary will desire milk for her coffee, so she will look in the fridge for milk; she will notice the date on the open container of milk; so, out of wariness of food poisoning, she will return that to the shelf and instead open the other container.

Let's assume (plausibly) that Mary does indeed go through the process just described. Now, there will exist a mapping φ between Joe's intermediate representations of Mary and her actual mental states. And, if we grant the Theory-Theorist that Joe has a passably accurate theory of Mary's mind, this mapping will preserve the normal progression N(x) of Joe's theory-of-mind inferential processes. Hence, if the Theory-Theorist is right, there will exist a system homomorphism from Joe's mentalizing procedure to Mary's mind. And hence, if the Theory-Theorist is right then this case would qualify as a case of mental-mental simulation. Indeed, this result is intuitive – Joe *does* simulate Mary insofar as he employs an inferential system that (quite roughly) mirrors the mental processes that Mary actually goes through.

However, notice that this is a very *abstractly-similar* form of simulation; for Joe's inferential process bears a homomorphism to Mary's mental process only at a very *coarse-grained* level of description. Clearly, this case involves a much more *abstract* similarity than the sort of similarity that Simulation Theorists typically envision. If Simulation Theory is to be interesting, it must not stake so weak a claim that even a clearly Theory-Theoretical case would entail its truth. I think the best way to avoid this risk of trivialization, while emphasizing what has been most central to most Simulation Theorists' views, is to strengthen Simulation Theory to a version that holds that mentalizing typically involves simulations that are quite *concretely*-similar.

> **Conclusion #1:** Many plausible versions of theory-theory will entail that multiple-step-mentalizations involve abstractly-similar simulation, though this is not a very interesting sense of simulation. If simulation theory is to stake an interesting claim, it must hold that *concretely-similar* simulations play a central role in mentalizing.[15]

7. Mentalizing Using Self as a Reference

I will now raise a second concern – namely, that a broad class of cases that intuitively *seem* to be clear cases of simulation will turn out *not* to involve simulation in the strong and interesting sense demanded by Conclusion #1.

Suppose Susie and Johnny are sitting in a classroom. The teacher asks Johnny, "What is 2+3?" How would Susie go about predicting what Johnny will say? The theory-theorist tells us that she would employ some sort of theory, and that this theory would somehow yield her prediction. Yet, intuitively, Susie doesn't need to employ much of a theory of Johnny's mental states to predict his answer – instead she need only *do the problem herself* and that gives her the prediction. On the surface, this certainly seems to be a case of simulation; Susie seems to be putting herself into Johnny's shoes, and the success of this venture seems to depend on her problem solving machinery being similar enough to his. Hence, it would seem, we should take this case to weigh in favor of Simulation Theory. I will argue, however, that this case is not nearly so clear-cut – indeed that her computation isn't actually a case of simulation at all!

Think about how exactly Susan goes about making this prediction. A crude, highly-intellectualized account might have it that Susie reasons as follows: "Johnny is a good student. Good students want to answer questions correctly. Johnny is smart enough to answer a simple math problem correctly. If Johnny wants to answer the question correctly, and is smart enough

[15] Goldman (in press) defines simulation in such a way that *all* cases of simulation are concretely similar, so he likely will not be opposed to Conclusion #1. However, I fear that his *general* theory of simulation may leave out far too many intuitive cases (e.g., computer simulations). I hope that the general theory I have presented here does well to capture our general intuitions, while still allowing (via Conclusion #1) a strong-enough-to-be-interesting formulation of Simulation Theory.

to answer the question correctly, then he will answer it correctly. So, I should predict that he'll say the correct answer. What is that? Let me see… 2+3? I'm pretty sure those add to 5. So, I predict that Johnny will say 'Five'." Of course, a more sophisticated account would hold that all of this occurs at a much more tacit level, and perhaps that it involves much different information-processing connections. But this crude picture will serve well enough for our purposes.

Now, the key question we need to answer is this: When Susie adds 2 and 3 together in her own head is she *simulating* the computation that Johnny is doing? I propose that the answer is no. Recall that, by our analysis above, a simulation must rely for its non-accidental success upon its reflecting significant aspects of the process being simulated. The success of Susie's prediction relies upon no such similarity between the way that she adds and the way that Johnny adds. For example, it might be that Johnny does the sum by extending two fingers, then extending three additional fingers, and then counting how many fingers he has extended. On the other hand (no pun intended), Susie may compute the sum by 'looking it up' on the addition-table she has memorized. There is no appropriate mapping between these two processes – neither can be used as a successful (concrete) simulation of the other. Susie's computation yields a perfectly good prediction of Johnny's answer. The success of this prediction is no accident, but it *does not rely* upon her computation mirroring his. As far as the non-accidental success of her prediction is concerned, it doesn't matter at all whether or not Susie's computation (concretely) mirrors Johnny's – all that matters is that they were both competent enough to arrive at the correct sum. Hence, Susie is not *simulating* Johnny's computation; instead she performs her computation *only* as a way to determine the correct answer to the problem – an answer that he might arrive at in an entirely different way.

To see this more clearly, consider a similar case. Let's imagine that Susie knows Johnny to be a math-prodigy, quite capable of doing complicated arithmetic in his head. Let's imagine that the question the teacher asks is "What is 127 times 259?" And, finally, let's suppose that Susie, herself, is quite incapable of doing such arithmetic in her head, but that she does have a pocket calculator handy. Susie would naturally predict Johnny's answer by entering the numbers into the calculator, and predicting that he will say whatever number the calculator spits out. I think it is intuitively clear that Susie's use of the calculator is *not* a case of simulation – she simply is not using the calculator to *simulate* the way that Johnny does the calculation, at least not in the strong and interesting sense of simulation that was demanded by Conclusion #1. Now, there is no important difference between this case and the 2+3 case. In both cases, Susie simply utilizes whatever means she has at her disposal to figure out what the correct answer is, and she predicts (by her theory) that Johnny will arrive at this answer too. It is intuitively clear that Susie is *not* simulating Johnny's computation when she uses the calculator. And there is nothing special about the 2+3 case that should make us say that adding 2 and 3 in her head *was* simulation, while using the calculator was not. Hence, we should conclude that neither of these computations is a case of (concrete, interesting) simulation.

I've been arguing just that Susie's *computation* is not a simulation of Johnny's *computation*. The simulation theorist still might claim that simulation plays a key role in some *other* part of Susie's mentalizing process. (E.g., it might be that Susie uses simulation to determine that Johnny will *desire* to answer the question correctly.) I will offer no argument here against *that* simulationist claim. My present purposes are satisfied just by pointing out that the 2+3 case *seemed* to be a strong intuitive case for simulation theory *because* it seemed clear

that Susie's computation was a simulation of Johnny's. Yet on closer inspection it turned out that this case doesn't provide *that* sort of intuitive support to Simulation Theory, after all.

This isn't limited just to mathematical cases. The simulation theorist, Jane Heal, takes a central case of mental-mental simulation to be the case where a mentalizer predicts some beliefs that a target will have on the basis of knowing other beliefs that target has:

> The nub of the simulationist proposal is that in thinking of another's thought what we do is take the subject matter of that thought, whether we believe in it or not, and think directly about it. The simulator and simulated are thus alike in that they think about the same subject matter, exercise the same intellectual capacities of understanding it and (if all goes smoothly) *move through the same reasoning processes* to the same conclusions about it. (Heal 1995, p. 35, my emphasis)

What makes this a *simulationist* proposal is the clause I have emphasized, which demands that non-accidentally successful simulations involve a concrete step-by-step similarity between the reasoning process of the mentalizer and that of her target. But, as any logic teacher knows, there are often many different lines of reasoning that would lead from a set of premises to a particular conclusion. If you trust that your target will, somehow or other, correctly determine whether a certain claim follows from certain premises, then an effective mentalizing strategy will be to take whatever route you can find from those premises to a conclusion about the truth of that claim. But this effective strategy does not involve concrete simulation, because it does not demand, for its non-accidental success, that the particular way in which you reach the conclusion be the same as the way in which your target reaches it. Once again, what initially seemed to be a clear-cut, interesting case of simulation turns out not to be.

Parallel considerations arise for cases of predicting *behavior* on the basis of known beliefs and desires. So long as you trust that your target is rational enough to choose the best means for achieving her ends, then you will do well to figure out what that best means would be, and predict that this is what she will do. This strategy is also not an interesting concrete case of

simulation, because it too does not depend for its non-accidental success upon there being a concrete similarity in the particular ways in which you and your target arrive at the conclusion that a particular means would be best.

Similar considerations also yield a plausible account of a puzzling empirical case: people's 'ego-centric empathy gaps' surrounding the endowment effect.[16] Subjects who own mugs tend systematically to *overestimate* how much other people might be willing to spend on a mug; while subjects who don't own mugs tend systematically to *underestimate* how much money mug owners might sell their mugs for. (Van Boven, Dunning and Loewenstein, 2000). One plausible interpretation of the endowment effect is that having a mug makes a person likely to suppose that a mug is *actually* worth more than non-mug-owners suppose that it is worth. This would plausibly account for subjects' poor predictions, so long as we assume that subjects predict a buyer's (or seller's) offer on the basis of how much the subject thinks a mug is *actually* worth. (This hypothesis is consistent with the claim that subjects do this by employing a crude theory.) According to this hypothesis, a subject does *make reference to* her own mug-valuation-judgments as she predicts what someone else would buy (or sell) a mug for. However, this self-reference *does not* constitute an interesting case of simulation, for the non-accidental success of such predictions in no way requires that the predictor go through anything (concretely) like the valuating process that the target goes through. Van Boven et al report that, in their statistical analysis, once the effects of introspected selling price on offers by buyers' agents[17] was taken

[16] Goldman (in press) considers ways in which this case might be accounted for by Simulation Theorists.

[17] In this experiment, subjects were divided into two groups. Each member of one group was given a mug to keep, while each member of the other group was shown, but not given, an identical mug. Experimenters then asked each subject to estimate how much they would be willing to sell such a mug for, were they to own a mug and have the opportunity to sell it. Van Boven et al. call this report the introspected selling price. Then, subjects were placed in the role of buyer's agent – they were given one chance to make an offer on a cup. If this offer meets or exceeds the minimum selling price of a randomly selected seller from a previous, similar, experiment then the subject would receive $10 minus the offer they made. On the other hand, if the offer is lower than the minimum selling price, then the subject gets nothing.

into consideration, "there was a substantial reduction in the variance accounted for by mug ownership, relative to the variance accounted for when mug ownership was the single predictor of participants' offers" (pg 72). In other words, mug-ownership led to less of an increase in subjects' final offers (as buyer's agents) in occasions where it led to less of an increase in how much subjects thought they *themselves* would be willing to sell a mug for ('introspected selling price'). This is entirely consistent with the hypothesis that *both* of these variables (introspected selling price and price offered as buyer's agent) depend critically upon how much the subject thinks a mug is *actually* worth.

I have argued that a number of cases that might seem to be strong examples of simulation turn out, under closer inspection, not to involve (interesting) simulation after all. One might worry that my argument plays upon technicalities in my definition of simulation. We may allay this worry by laying the technicalities aside and phrasing my argument more intuitively. The concern is that there is a large class of mentalizing cases in which the mentalizing agent does *not* initiate a process that corresponds step-by-step (concretely) to the target's mental process. Instead, it seems plausible to think that mentalizers very often go through *quite different* step-by-step processes from their targets. Mentalizing is often successful for good reason. But that reason, in the cases of concern, is *not* because of any concrete step-by-step similarities between the involved processes, but *just* because the involved processes are processes that tend to arrive at similar conclusions. This is consistent with both Theory-Theory and those uninteresting forms of Simulation Theory that allow that mental-mental simulation might involve very abstract similarities. The concern is, however, that *interesting* versions of Simulation Theory – versions that hold that mentalizers go through processes that closely correspond step-by-step to their targets – will fail to account for these problematic cases.

> **Conclusion #2:** Many cases that seem, intuitively, to support Simulation Theory turn out to be such that, at the intuitively critical juncture, what seemed to be simulation turns out to be only a very abstract input-output simulation, at best. Given Conclusion #1, this sort of simulation will not support any interesting formulation of Simulation Theory.

8. Possible Simulationist Responses.

Even if many cases have the problematic character described above, there may still be many interesting cases of concrete mental-mental simulation. The most promising cases for the Simulation Theorist will be ones where different people would *not* be likely to reach the same conclusion by quite independent avenues – e.g., cases that involve dedicated systems that all people share and/or cases where *what-conclusion-you-arrive-at* is dependent upon the subtle foibles of human mental machinery. Particular examples include emotional and visceral responses, dedicated perceptual or motor systems (e.g. optical illusions, garden path sentences), open-ended questions (e.g., "How would he react to this?"), speed-intensive tasks (e.g., quickly estimating math problems), and the prediction of errors or mistakes.

For all I've said here, Simulation Theory may be the correct account of many of the cases just listed. But what is a Simulation Theorist to say about the problematic cases like those presented in the preceding section – cases where non-accidental success does not require a concrete step-by-step correspondence between the two processes?

One possible response would be to seek a *broader* definition of simulation than the one I presented. However, coming up with a definition that fits the bill may not be easy. Authors such as Goldman and Gordon who suggest that successful simulation involves some sort of 'isomorphism' will be hard-pressed to find an interesting isomorphism between the diverse avenues by which one might do arithmetic (or determine a rational course of action, or judge the value of a mug). Similar problems will face *anyone* who accepts the intuitive premise that *if* process P_1 is a successful simulation of process P_2 *then* P_1 must reflect in *some* interesting way

what is going on in P_2. I doubt that any definition that is broad enough to accommodate these problematic cases would satisfy our intuitive conception of what counts as *simulation*.

But this is not to say that there aren't plausible accounts of mentalizing along similar lines to what '*Simulation Theorists*' have suggested. I'm arguing only that so-called 'Simulation Theorists' have not, in all cases, been talking about *simulations*. Let us consider three more-or-less simulation-like alternatives that traditional 'Simulation Theorists' might move to defend in the place of full-fledged concrete mental-mental simulation.

First, why not hold just that many cases of mentalizing involve having the mentalizer actually enter into brain states that closely correspond to the target's? Why care whether this correspondence is *purposeful* or merely an accident? This move accounts for the (perhaps rare) cases in which Susie *does* add 2 and 3 together in just the same way as Johnny does. However, it does nothing to account for why many people would predict Johnny's answer correctly *even if* their brain states did not closely track his. Insofar as such cases constitute a broad class of phenomena, the present suggestion will fail to provide a general theory of mentalizing.

Second, why not hold just that a significant number of mentalizing cases involve feeding pretend inputs through an off-line device? Why care whether these qualify, technically, as cases of simulation? This attitude nicely fits one intuitive presentation of Simulation Theory as the view that, when someone mentalizes, she takes some useful device of her own off-line, feeds pretend inputs into it, and uses its outputs to draw conclusions about her target. The key claim in this intuitive picture is just that cases of mentalizing involve running pretend inputs through an off-line device, and *not* that this device concretely mirrors the functioning of some device in the target's head.

This approach is certainly attractive, but I am concerned with the word 'pretend'. I don't know what it would mean to say Susie is only *pretend-adding* while Johnny is *really-adding*,[18] or that Susie is *pretending* to judge the value of a mug, while Johnny is *really* judging its value.[19] Hence, I think we shouldn't construe the problematic cases above as involving off-line *pretense*. Instead, I recommend a third option that retains much of the attractiveness of the second, but without the problematic notion of *pretense*.

9. Mentalizing by Equi-Finality.

I propose that a significant class of mentalizing cases involve the employment of *equi-final devices*. Two devices are ***equi-final*** if, when operating successfully (with respect to their purposes), they will produce corresponding representations as output whenever they are fed corresponding representations (from a given class) as input.[20] For example, Susie's memorized addition table and Johnny's finger-counting are *equi-final* with respect to the class of small positive integers. We may define ***Mentalizing-by-Equi-Finality*** as purposefully employing a mechanism in a mentalizing process, where the non-accidental success of this purpose requires (1) that the employed mechanism be *equi-final* to some mechanism employed by the target, (2) that these two mechanisms are fed corresponding representations as input, and (3) that the two mechanisms both successfully produce their output in accordance with their ordinary purposes.

---

[18] One might hold that Susie is *pretending-to-be-Johnny-adding*. However, it is not at all clear to me that she is doing that at all! Indeed, she might even know that he adds on his fingers, but not care to go through that rigmarole. I see no reason to suppose that her doing the calculation her own way is somehow *pretending-to-be-Johnny*. Instead, she calculates the sum, just so that *she'll* know what the right answer is.

[19] The case of determining a rational course of action may be *more* plausible as a case of pretense. But even in that case, I'm not convinced that the mentalizer *needs to* pretend to be someone else. Instead, she might just ask herself questions like "How might someone achieve X?" Certainly, this would exercise mechanisms that she uses in her own reasoning, but, again, it's not clear that this is a *pretend-exercise*.

[20] Defined this way, an equi-final mechanism may be employed as an (uninteresting) input-output simulation. However, this definition of equi-finality relies upon having some *independent* definition of what it is for two representations to correspond, while my definition of simulation above defined an appropriate mapping in terms of preserving the similarity in causal structure of the two systems.

This equi-finality account has several theoretical advantages over an account in terms of mental-mental simulation. Most importantly, the equi-finality approach accounts for a mentalizer's non-accidental success in the problematic cases above – cases where mentalizer and target might use quite different avenues to reach the same conclusion. Also, the equi-finality approach is less tied to the idea of pretense than is traditional Simulation Theory. Equi-final devices (at least the ones employed in the problematic cases) are calculating- or solution-finding-mechanisms. Insofar as it is unclear what it would mean to *pretend to* do a calculation or to *pretend to* find a solution, this move away from pretense is a virtue of the equi-finality approach.

On the other hand, mental-mental simulation allows that one might succeed non-accidentally by using an off-line device that *fails* to achieve its ordinary purpose. For example, a simulation may be successful *because* it involves the right sort of lapse in reasoning. The equi-finality approach handles these cases less easily, for Mentalizing-by-Equi-Finality succeeds non-accidentally only when both mechanisms *succeed* at their purposes. Mental-mental simulation also may provide a more plausible account of cases that involve several mechanisms and/or mechanisms that do not have a specific representation as their normal output (e.g., perhaps, emotional mechanisms).

These complementary weaknesses and advantages suggest that Mental-Mental Simulation and Mentalizing-by-Equi-Finality might *both* be tools worth employing in the modern Hybrid-Theorist's toolbox.[21] Both involve the employment for mentalizing of devices that see regular 'on-line' use. And, both allow that significant portions of mentalizing tasks are

[21] I think that the most plausible hybrid theory's box of mentalizing tools must also include large roles for broadly theory-theoretic reasoning, and also for broadly associative recognitional capacities. Given the probable ubiquity of our use of these other tools, I think it was antecedently improbable that simulation played all that large of a role in human mentalizing, even prior to the arguments given here. My arguments here serve to trim the role for simulation even further, and to introduce a previously unrecognized tool – equi-final devices – to our known supply of useful mentalizing tools.

performed by non-theory-of-mind devices. Indeed, I think that many traditional 'simulation theorists' have (mistakenly) claimed equi-final devices as cases of simulation. I've argued here that, technically and intuitively, many of these devices are not employed as simulations. However, that certainly does not entail that the enterprising Hybrid Theorist can't claim equi-final devices as a phenomenon that bears significant similarity to full-blown cases of simulation. In this way, the Hybrid Theorist might retain much the motivating insights behind Simulation Theory, but in a more robust and plausible hybrid view.[22]

[22] I owe a special debt of gratitude to Alvin Goldman. Taking a seminar with him, reading an early manuscript of his book in preparation, and discussing these topics with him have all influenced my thinking on these matters in countless ways. I am also indebted to helpful comments from Stephen Biggs, Shaun Nichols, Sarah Wright, and audiences at the University of Arizona, the Society for Exact Philosophy and the Society for Philosophy and Psychology.

References.

Davies, M. and T. Stone. 1995. "Introduction." *Folk Psychology: The Theory of Mind Debate*. Ed. Martin Davies and Tony Stone. Cambridge: Blackwell. 1-44.

Goldman, Alvin. 1995. "Interpretation Psychologized." In *Folk Psychology: The Theory of Mind Debate*. Ed. Martin Davies and Tony Stone. Cambridge: Blackwell. 74-99.

_____. 2003. "Conceptual Clarification and Empirical Defense of the Simulation Theory of Mindreading", in C. Kanzian, J. Quitterer, and E. Runggaldier, eds., *Persons: An Interdisciplinary Approach*. Vienna: Springer.

_____. (in press). *Simulating Minds: The Philosophy, Psychology and Neuroscience of Mindreading*. Oxford UP.

Gordon, Robert. M. 1995. "The Simulation Theory: Objections and Misconceptions." In *Folk Psychology: The Theory of Mind Debate*. Ed. Martin Davies and Tony Stone. Cambridge: Blackwell. 100-122.

Heal, Jane. 1995. "How to Think About Thinking." In *Mental Simulation*. Ed. Martin Davies and Tony Stone. Cambridge: Blackwell: pp. 33-52.

Lewis, D. 1984. "Putnam's Paradox", Australasian Journal of Philosophy 62: 221-236.

Millikan, R.G. 1984. *Language, Thought, and Other Biological Categories.* Cambridge: MIT Press.

Putnam, H. 1977. "Models and Reality." In *Realism and Reason*. Cambridge: Cambridge University Press, 1983, pp. 1- 25.

Van Boven, L., D. Dunning and G. Loewenstein. 2000. "Egocentric Empathy Gaps Between Owners and Buyers: Misperceptions of the Endowment Effect." Journal of Personality and Social Psychology. Vol 79. No. 1. pp. 66-76.