Proceedings of the 5th International Conference on Interdisciplinary Humanities and Communication Studies
DOI: 10.54254/2753-7064/2026.39029

Cognitive Instability of Mindreading: Why Simulation Theory Cannot Support Process Reliabilism

Jingnan Lin

Department of Philosophy, School of Philosophy and Social Development, Huaqiao University, Xiamen, China
2427111024@stu.hqu.edu.cn

Abstract. Alvin Goldman's simulation theory offers a compelling description of how human beings read other minds, yet, insofar as it is advanced as a cognitive practice that yields knowledge, it must answer to epistemological standards, most pointedly to Goldman's own process reliabilism, on which a true belief counts as knowledge only if it is produced by a cognitive process that reliably leads to truth. This paper argues that the simulation mechanism lacks any independent error-correction procedure and therefore cannot screen out contamination arising from the subject's own states, such as biases and emotions. This structural deficit makes the accuracy of simulation heavily dependent on the contingent factor of how similar the target is to the subject, so that its truth-production rate fluctuates sharply across contexts and fails reliabilism's demand for cross-contextual stability, leaving it trapped in what Pritchard calls veritic luck. We do not deny the descriptive value of simulation theory; we contend only that, until simulation is finely subdivided and its operative environments delimited, the assertion that simulation produces knowledge cannot satisfy the reliabilist standard. The argument is strictly normative in its target: the paper does not dispute that simulation plausibly describes how mindreading is in fact carried out, but only whether, given the structural deficit identified below, the outputs of simulation are yet entitled to the status of knowledge.

Keywords: simulation theory, process reliabilism, epistemic luck, veritic luck, environmental sensitivity

1. Introduction

1.1. Background: simulation theory as a descriptive theory of mindreading mechanisms

In the overlapping domains of contemporary philosophy of mind and cognitive science, Alvin Goldman systematically developed and defended simulation theory in Simulating Minds, making an important contribution to the explanation of human mindreading [1]. As a descriptive theory, simulation theory holds that people do not rely on implicit psychological laws, as the "theory theory" claims, but instead predict and explain others' behaviour and mental states by generating "pretend beliefs" and "pretend desires" within their own cognitive systems and thereby taking up the target's point of view [1].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7064/2026.39029

This paper does not dispute the explanatory power of simulation theory at the descriptive level. Descriptive power, however, does not automatically translate into epistemological legitimacy. When simulation theory is elevated into a theory about knowledge, that is, when it claims that the simulation process can generate genuine knowledge rather than mere true belief, it must confront the central question of epistemology: what standard must a belief-forming process satisfy in order to produce knowledge?

1.2. Why reliabilist standards are needed to evaluate simulation theory

It is worth noting that subjecting simulation theory to epistemological scrutiny is not an "external" critique imposed from outside Goldman's own framework. Quite the opposite: Goldman himself is a principal champion of process reliabilism. In his foundational 1979 paper "What Is Justified Belief?" he proposed that a belief is justified only if it is produced by a reliable cognitive process [2], and he carried this framework through his 1986 monograph Epistemology and Cognition, where a true belief counts as knowledge only if it results from a process that reliably leads to truth [3]. Any cognitive practice that claims to yield knowledge must therefore possess enough stability to keep producing true beliefs across relevant environments.

Evaluating simulation theory within Goldman's own reliabilist framework is thus not only legitimate but a necessary step for testing the internal consistency of his theory. If simulation really can produce knowledge, it must pass the reliabilist test. The epistemological payoff of this internal strategy is twofold. It raises the standard of argument from an external imposition, which a simulation theorist could deflect by rejecting reliabilist criteria wholesale, to an internal test of consistency that Goldman's own commitments authorise. It also relocates the burden of proof: if process reliabilism is defended as the correct account of knowledge, then any candidate source of knowledge, simulation included, must be shown to satisfy it, so the disagreement is no longer about which standard is to be applied but about whether one independently motivated standard is in fact met.

1.3. This paper's thesis and structure

The core thesis of this paper is that the simulation mechanism lacks an independent error-correction procedure and so cannot effectively screen out the contamination of the simulation process by the subject's own states, biases, emotions, and prior beliefs. This structural defect makes the accuracy of simulation highly dependent on the contingent factor of the similarity between the target and the subject; its truth-production rate fluctuates sharply across situations and fails reliabilism's requirement of cross-contextual stability, and it therefore falls into the dilemma of "epistemic luck" in Pritchard's sense [4]. Part II establishes the evaluative standard; Part III argues for simulation's structural defect; Parts IV and V develop two consequences of that defect and respond to the "rough reliability" defence; Part VI concludes.

2. Establishing the evaluative standard: reliabilism and veritic luck

2.1. Goldman's process reliabilism

The core claim of Goldman's process reliabilism can be summarised as follows: a belief B is justified if and only if B is produced by a reliable cognitive process [2]; a process is "reliable" in so far as it tends to produce true rather than false beliefs, that is, it enjoys a sufficiently high truth-

DOI: 10.54254/2753-7064/2026.39029

production rate. In "What Is Justified Belief?" Goldman rules out foundationalism and coherentism and proposes a historical-reliabilist account: the justification of a belief depends not only on the belief itself but also on the causal history of the process that produced it [2]. He also stresses the problem of "process types", "vision" and "vision at thirty yards in dim light" are two types whose reliability may differ [2], a problem that later became one of reliabilism's chief difficulties, the so-called generality problem [5].

In Epistemology and Cognition he further distinguishes perception, memory, and reasoning as different process types, and, for belief-dependent processes that take existing beliefs as inputs, introduces "conditional reliability" [3]: when the input beliefs are justified and the process operates reliably upon them, the output belief is justified as well [2]. This notion is crucial here, because simulation is itself a belief-dependent process: it takes the subject's information about the target's situation as input and, after a simulation operation, outputs a belief about the target's mental state.

2.2. Pritchard's concept of veritic luck

Pritchard, by contrast, holds that satisfying the reliability condition is not by itself sufficient for knowledge. In Epistemic Luck he argues for the necessity of an anti-luck requirement [4] and distinguishes "generalised luck" (a true belief formed by a reliable process whose truth nonetheless depends on the environment's being fortuitously normal) from "veritic luck" (where the luck lies in the process itself) [4]. His focus is veritic luck: a belief that is in fact true, yet false in most nearby possible worlds, owes its truth to luck rather than to cognitive ability. On this basis he proposes the safety condition, for a belief to constitute knowledge it must not easily have been false in nearby possible worlds [4]. In "Veritism and Epistemic Value" he adds that if knowledge were merely true belief formed by a reliable process, its value would be "swamped" by the value of the true belief itself, the reliable process figuring only as an instrumental means to truth [6]; resolving this swamping problem requires emphasising the extra value knowledge gains by excluding epistemic luck. A reliable process must therefore not only produce true belief in the actual world but remain stable across relevant possible worlds, precisely the challenge simulation theory faces. The contrast between the two species of luck should be stated precisely, since the argument that follows turns on it. Generalised luck is a defect of placement: a genuinely reliable process yields a true belief, yet only because the environment happens, by good fortune, to be of the kind in which that process works. Veritic luck is a defect of the process itself: the very mechanism that yields the truth is one that would easily have yielded a falsehood in nearby conditions, so the truth is attributable to the process only in name. The decisive question is thus whether the luck arises from a structural feature of the process or merely from an accidental handicap in its external environment, a distinction that will be applied directly to simulation in the next section.

2.3. Conceptual distinction between epistemic luck and environmental sensitivity

Placing simulation theory within the reliabilist framework requires a key distinction: which kind of environmental sensitivity constitutes a source of epistemic luck, and which is acceptable? Pritchard notes that every cognitive process is environmental-sensitive in some sense, vision depends on light, memory on the normal functioning of storage and retrieval, yet not all such sensitivity amounts to epistemic luck. The decisive question is whether the luck arises from a structural defect of the process itself or merely from an accidental handicap in the external environment [4]. This paper holds that simulation's problem belongs to the former. A process equipped with objective correction mechanisms that detect and amend errors caused by environmental change has manageable, non-

DOI: 10.54254/2753-7064/2026.39029

lucky sensitivity, vision, for instance, possesses calibration mechanisms such as binocular disparity. Conversely, a process that lacks an independent error-correction procedure, whose truth-production rate depends entirely on the accidental match between environment and process, bears the structural signature of epistemic luck. This is the core of our critique of simulation theory. Even if Goldman's various "special-environment" versions of reliabilism, "non-manipulative environments," the "normal world", are adopted, as Graham has reconstructed them [7], simulation still confronts a fundamental reliability challenge.

3. The structural defects of simulation theory: absence of an independent error-correction procedure

3.1. Internal vulnerability of the simulation process: intrusion of real states

As argued above, a reliable process must retain stable output in the face of environmental change. Yet an examination of simulation's core mechanism, "off-line simulation", reveals exactly the opposite: it lacks the isolation such stability requires. Simulation theory holds that the subject generates pretend beliefs and desires about the target's mental states and feeds them into her own cognitive system to understand the other [8]; the crucial requirement is "isolation", the pretend states must not be contaminated by the subject's genuine beliefs and desires. Goldman himself, however, concedes in "In Defense of the Simulation Theory" that "real states may infect simulations in which they should not appear" [8] , that is, isolation failure. Wallin points out that simulation theorists often explain the "egocentric bias", over-attributing one's own states to others , precisely through such isolation failure [9], and she asks: if simulation fails to isolate even in simple adult tasks, how can we trust its reliability in more complex settings [9]? Keysar and colleagues corroborate this: when an informed speaker knows a bag contains tape while the addressee does not, yet both see the bag, the speaker still involuntarily orients toward the bag when the addressee mentions "tape" [9] , adults cannot fully suppress their own perspective. If simulation is contaminated by the subject's own knowledge in so simple a task, its reliability is all the harder to defend in situations of ambiguous cues, incomplete information, or a subject under stress or fatigue.

Two features of this concession deserve to be made explicit. First, Goldman's admission is one of principle rather than of degree: he grants that isolation is a requirement of the account and that its violation is possible, yet he treats contamination as a local and correctable impurity rather than as a systematic tendency, and the remedies he sketches, most notably the restriction of simulation to domains in which subject and target share a common cognitive style, amount to narrowing the scope of the theory rather than repairing the mechanism. Second, Keysar's bag-and-tape result should be read as more than evidence that adults fail to suppress their own perspective. Since the interference manifests both where the simulation is constructed and where its output is expressed in an orienting response, it indicates that neither the input nor the output channel of the mechanism is sealed against the subject's own knowledge. On this reading egocentric bias is not an occasional leak in an otherwise closed process, but a symptom of a process that has no door available to close.

3.2. The indiscernibility of "pretend states" and real states at the phenomenological level

That isolation failure is so frequent has a deeper phenomenological root: "pretend mental states" and "real mental states" lack a clear experiential boundary, so the subject is left without metacognitive monitoring capable of detecting whether a simulation has been contaminated. Tooming accordingly raises a systematic objection to Goldman: simulation theory depends on the core concept of a

DOI: 10.54254/2753-7064/2026.39029

"pretend mental state," yet pretend and real states are not phenomenologically clearly bounded [10], when we imagine another's feeling, the imagining itself activates neural and physiological mechanisms tied to genuine experience, blurring the boundary. This indiscernibility has an important epistemological consequence: if the subject cannot tell whether an output rests on a clean or a contaminated simulation, the process lacks an intrinsic error-detection mechanism. Perception may also err, but it possesses independent correction cues, consistency checks, cross-modal verification, expectancy violation, that simulation lacks.

This indiscernibility is not a mere conceptual stipulation; it is borne out by empirical work on the mechanisms of simulation. Neuroimaging studies of action observation report that observing another's movement recruits the observer's own motor representations via the mirror-neuron system, which is why simulated and executed actions compete for shared computational resources. Affective simulation shows a comparable overlap: imagining another's pain or disgust elicits autonomic and facial-muscle responses of the same kind as experiencing it, though smaller in magnitude, and participants can be experimentally misled about the source of such arousal. The epistemological lesson is direct. Where the signature of simulating an emotion is a weaker token of the very same state, no internal marker remains by which the subject can tell that her simulated feeling has been seeded by her own mood, so the absence of an error-detection mechanism is not a design oversight but a consequence of the shared neural substrate.

3.3. Locating the structural defect: absence of an independent filter

Combining the empirical isolation failure of 3.1 with the phenomenological blurring of 3.2, simulation's structural defect can be located as follows: the simulation mechanism lacks an objective, independent "error-correction procedure" that could remove contamination from the subject's own states. Reliabilism requires a process not merely to produce true belief in typical environments but to possess intrinsic stability, a comparatively consistent truth-production rate across changing inputs. Perception is partly reliable precisely because it enjoys independent world-anchoring: binocular disparity calibrates depth, familiarity corrects illusion, cross-modal verification detects error. Simulation is different: taking the subject's information about the target's situation as input, it feeds pretend states into the subject's own system and outputs a belief about the other's mind. Two problems follow. First, the input itself may be unreliable, already laced with misinformation or egocentric bias; second, simulation lacks an independent filter to detect whether it has been contaminated by real states. As a paradigmatic belief-dependent process, its output is wholly at the mercy of the input's reliability and the process's conditional reliability; once the input is unreliable and hard to check, the simulation is compromised. Reasoning, by contrast, can be audited by logical rules, memory by other memory sources; simulation alone depends on the subject's own cognitive system, which cannot be corrected from outside when the system itself is biased. Its reliability is therefore a "dependent reliability", contingent on accurate inputs and favourable conditions, lacking an independent anchor, which Grundmann has analysed as rendering it peculiarly fragile under input variation [11].

DOI: 10.54254/2753-7064/2026.39029

4. Consequence one: contingent dependence on "similarity"

4.1. The conditional nature of simulation accuracy: the degree of similarity between others and self

The structural defect yields a direct consequence: the accuracy of simulation depends heavily on the contingent factor of how similar the target is to the subject. Simulation's basic logic, placing oneself in the other's situation, generating a pretend state, and observing one's own system's output, presupposes that "the other and I are similar in the relevant respects." Where character, values, and emotional-response patterns converge, the output may be accurate; where they diverge, the same mechanism systematically errs. Wallin's analysis above shows that simulation is prone to isolation failure; if the isolation mechanism easily fails, then the truth-production rate will fluctuate with the variable of "similarity": when similarity is high, self-projection may coincidentally hit the truth even under imperfect isolation; when similarity is low, across cultures, temperaments, or moods, the very mechanism deviates. Simulation is thus not a tool that can be freely calibrated to the target, but a process limited by its own structure and easily tainted by egocentric bias.

4.2. The contingency of "environmental match": the initial appearance of luck

That the truth-production rate fluctuates so violently with similarity is exactly what reliabilism guards against, a reliable process should be stable across contexts, not depend on a lucky "environmental match." Consider a weather forecaster who integrates barometric pressure, temperature, and humidity: though he occasionally misses, we still deem his process reliable because it commands multiple sources of evidence and correction mechanisms, and its output does not swing wildly with slight changes in the weather system. Simulation is the opposite: its truth-production rate hangs on a single dimension, how similar the target is to the subject, coincidentally hitting under high similarity and systematically missing under low. Such fluctuation is no accidental error but the inevitable result of a mechanism that takes the self as its model: a model's applicability naturally depends on how well it matches its target. This contingent dependence on "environmental match" already displays the signature of epistemic luck: the truth of the belief no longer stably derives from the reliable operation of the process, but depends on contingent factors beyond the subject's full control.

The comparison is sharper than it first appears, because what does the argumentative work here is not the number of data sources but the stability they secure. The forecaster is creditworthy because his output is insensitive to the deterioration of any single input: when one variable degrades, the remaining channels hold the verdict in place, and this is precisely the cross-situational constancy that reliabilism demands of a process that is to yield knowledge. Simulation carries only one parameter, the degree of similarity between subject and target, so a change along that single dimension propagates directly and unbuffered into the truth-value of its output. Stability across contexts, rather than average accuracy within one, is the criterion reliabilism imposes, and it is the one dimension on which the simulation mechanism is structurally deficient.

DOI: 10.54254/2753-7064/2026.39029

5. Consequence two: high environmental sensitivity of true belief production

5.1. The impact of input-end contamination on output

If Part IV exposed simulation's contingent dependence on the external variable of similarity, this part goes further: even when similarity is high, the internal information-processing of simulation is itself highly fragile. As a belief-dependent process, simulation takes the subject's information about the target's situation as input and outputs a belief about the other's mental state after an operation [11]; this structure makes it acutely sensitive to input contamination. First, if the subject's perception or memory of the target's situation is mistaken, the error is fed directly into the simulation and yields a false output, while the simulation possesses no independent means of checking the input's truth, it can only process the data it is given, not cross-verify across modalities as perception does. Second, as Wallin analyses, the subject's prior beliefs and emotions may contaminate the simulation through isolation failure [9]: a subject holding a strong negative bias toward another may overestimate the other's hostility; this contamination is no accident but follows from simulation's having to use the subject's own, non-isolable cognitive resources. Third, the "off-line" character of simulation means it receives no immediate corrective feedback from the outside: the subject feeds pretend states into a closed system and observes the output, and without an independent external source of verification she cannot determine whether that output genuinely tracks the target's actual state.

5.2. Responding to the "rough reliability" defence

Facing these objections, the simulation theorist may appeal to "rough reliability": even if simulation is not perfectly reliable, so long as it yields roughly correct attributions in everyday situations, it satisfies the reliability required for knowledge. Goldman himself has a comparable strategy: as Graham reconstructs it, in response to counterexamples such as Norman the clairvoyant he progressively introduced limits like "normal environments" and the "normal world," confining the assessment of reliability to the range in which the process functions normally [7]. Yet this defence does not suffice for simulation. First, "rough reliability" needs a concrete threshold: set too low (say 51%), plainly unreliable processes such as coin-flipping qualify, violating reliabilist intuition; set too high, then given the pervasiveness and stubbornness of egocentric bias, simulation may simply not pass. Second, as Graham notes, normal-circumstances reliabilism faces the difficulty of delimiting the "normal environment" [7]: if "normal" is defined as whatever environment makes simulation reliable, the account turns viciously circular, one must first independently fix which environments are normal in order to assess reliability, yet reliability is precisely what varies with environment. Third, and most crucially, "rough reliability" ignores directional bias: reliabilism cares not only about the quantity of the truth-production rate but about the type of error. A process that biases systematically in one direction, simulation systematically overestimating the similarity between other and self, retains a character of systematic unreliability even if its average accuracy meets the threshold: it is not "broadly reliable, occasionally mistaken" but "doomed to fail under specific conditions." Absent an independent error-correction mechanism, then, simulation's environmental sensitivity is systematic rather than accidental and cannot be dissipated by the "rough reliability" defence.

The circularity deserves a closer look, since in the present case it is not merely formal but operational. To assess whether simulation is reliable in normal environments, one must first determine, independently of any simulation, which environments count as normal. Yet the access ordinary subjects have to the very features that make an environment cognitively hospitable, the

DOI: 10.54254/2753-7064/2026.39029

similarity of one's interlocutor, the transparency of the cues, the absence of stress or fatigue, is itself largely a matter of simulated attribution. The procedure is therefore required to deliver its own presupposition: it can certify reliability only after it has assumed that the capacity under audit is already functioning correctly. The assessment of simulation's reliability cannot so much as get under way, and the appeal to normal environments consequently rescues reliabilism in general while leaving this particular process unevaluated.

5.3. From environmental sensitivity to veritic luck: a complete argument

The analysis permits a complete argument from simulation's environmental sensitivity to veritic luck. First, the simulation process lacks an independent error-correction procedure, so its reliability depends entirely on the accuracy of its inputs and on the similarity between subject and target [11]. Second, because egocentric bias is pervasive, simulation's truth-production rate fluctuates violently across subject–target pairings, it may be right under high similarity and systematically wrong under low [9]. Third, this fluctuation is no accidental error but the inevitable upshot of a mechanism that takes the self as its model, where the degree of model–target match is a variable the subject cannot fully control. Fourth, by Pritchard's safety condition, a belief constitutes knowledge only if it remains stable in nearby possible worlds and is not true merely through fortuitously favourable environments [4]; the beliefs simulation generates fail precisely this condition: in a slightly different environment, a target very unlike the subject, the same simulation yields a different, even false, output. Priest adds that safety should be understood as a modal notion, concerning stability across nearby worlds rather than probability in the actual world [12]; simulation's problem is not that its probability of truth is low in the actual world but that its structure cannot guarantee stability in nearby ones. Rodríguez further notes that not only process reliabilism but also sensitivity and safety theories face the generality problem of fixing the type at which reliability is assessed [5], a problem especially acute for simulation. Should we classify the process as "mental simulation" or " simulation laden with egocentric bias"? The former seems fairly reliable, the latter far less; and the choice depends on whether we have already independently assessed simulation's reliability, which is circular. Simulation's beliefs therefore cannot meet reliabilism's strict demand.

It follows that the necessity established in the third premise is not confined to that premise but carries through to the conclusion. The instability just described is not the contingent effect of an unfavourable environment, which a change of circumstances might repair, but the inevitable upshot of a mechanism that takes the self as its model: for so long as the model is the subject's own cognitive apparatus, the degree of model-target match remains a variable the subject cannot settle, and every nearby world in which that variable shifts is a world in which the output shifts with it. The modal fragility of simulation is thus inherited from its architecture, and it is this inheritance, rather than any observed error rate, that forfeits its claim to knowledge.

6. Conclusion

6.1. Summary of this paper's core contributions

Beginning from Goldman's own process reliabilism, this paper has examined the epistemological standing of simulation theory and concluded that its qualification as a source of knowledge does not hold. Because simulation must borrow the subject's own cognitive resources, whose biases and emotions cannot be fully isolated, the mechanism lacks an objective, independent error-correction procedure; consequently its accuracy hangs on the contingent variable of "similarity," its truth-

DOI: 10.54254/2753-7064/2026.39029

production rate swings across contexts, and even retreating to "rough reliability" cannot dissolve this systematic, directional bias, so that the beliefs it generates lose stability in nearby possible worlds, that is, they are hostage to epistemic luck. The upshot is a single line of reasoning: structural defect → environmental sensitivity → veritic luck. The reach of this line of reasoning, however, extends beyond simulation theory. Its three links, an error-detection procedure that cannot operate independently of the mechanism it is meant to monitor, an output governed by a single unbuffered parameter, and the resulting modal fragility, together specify a general test that any self-modelling cognitive mechanism, introspection, empathy, imaginative rehearsal or embodied interpretation alike, must pass before its deliverances may be credited as knowledge. Applied to the present case, the test returns a single verdict, and it is this that the preceding sections were designed to establish with precision: the value of simulation theory lies in description, while its difficulty lies in norm.

6.2. Directions for future research

It must be stressed that this critique does not deny simulation theory's descriptive value; it separates that value from the theory's normative claims. The work genuinely still to be done is threefold: to give "simulation" a finer-grained typology, distinguishing which sub-processes, for instance those grounded in inferential reasoning, carry their own correction mechanisms from those, such as purely affective resonance, that do not; and to delimit strictly the "relevant environments" in which simulation is effective, perhaps within highly homogeneous communities where subject and target barely differ. A third direction, which we raise as a genuinely open question rather than as an objection, concerns the trainability of the simulation mechanism. If subjects can, through learning, reflection or metacognitive training, sharpen their awareness of their own biases and emotional intrusions, then the absence of an innate error-detection mechanism need not be fatal: a learned and externally scaffolded calibration might partially compensate for the structural deficit. Whether it does so is a question that empirical and conceptual work can address jointly, the former testing whether debiasing or perspective-training protocols measurably improve the accuracy of attribution, the latter asking whether trained monitoring constitutes a genuine though non-independent error-preventive mechanism in the reliabilist sense. This avenue matters because it would offer simulation theory a more hopeful epistemological prospect than the two options already canvassed, environmental delimitation and typological refinement. Only after these three steps could simulation's reliability be re-assessed.

6.3. Theoretical prospects

Until such refinement is achieved, simply asserting that "simulation produces knowledge" is hard to sustain under reliabilist scrutiny. This is not a call to abandon simulation theory but a prompt to state its epistemological claims more cautiously, distinguishing its descriptive value from its normative ambition, and re-evaluating its eligibility as a source of knowledge only under well-specified conditions. The critique thus exposes a deeper tension: holding simulation theory to be simultaneously a psychological theory of mindreading and an epistemological theory of knowledge. As psychology it may describe whatever strategy humans in fact use, however imperfect epistemically; as an epistemological theory it must satisfy the exacting standards of reliability and safety, a line that ought to be drawn with care. Stated in its strongest form, the difficulty diagnosed here is therefore not a technical defect that further research might patch, but a structural one, arising from the theory's attempt to discharge two tasks of different logical type at once. As a psychological theory, simulation theory answers a descriptive question, namely how human beings in fact come to

DOI: 10.54254/2753-7064/2026.39029

attribute mental states to one another, and its adequacy is measured by empirical fit. As an epistemological theory, it answers a normative question, namely by what right such attributions claim the status of knowledge, and its adequacy is measured by reliability and safety. These two standards are independent of one another: a mechanism may describe the actual strategy of cognition with perfect fidelity and still fail to satisfy the conditions under which its outputs count as knowledge. The separation, and the frequent interweaving, of the two standards is the deeper anchor of the whole argument, and acknowledging it is what permits simulation theory to be retained at full strength in the first register while being withheld from the second.

References

[1] Goldman, A. I. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford University Press.
[2] Goldman, A. I. (1979). What is justified belief? In G. Pappas (Ed.), *Justification and knowledge* (pp. 1–23). Dordrecht:Reidel.
[3] Goldman, A. I. (1986). *Epistemology and cognition*. Harvard University Press.
[4] Pritchard, D. (2005). *Epistemic luck*. Oxford:Clarendon Press.
[5] Rodríguez, I. (2026). The generality problem for process reliabilism, sensitivity theory, and safety theory. *Synthese, 207*, 205.
[6] Pritchard, D. (2016). Veritism and epistemic value. In B. P. McLaughlin & H. Kornblith (Eds.), *Goldman and his critics* (pp. 199–218). Wiley.
[7] Graham, P. J. (2017). Normal circumstances reliabilism: Goldman on reliability and justified belief. *Philosophical Topics, 45*(1), 33–62.
[8] Goldman, A. I. (1992). In defense of the simulation theory. *Mind & Language, 7*(1–2), 104–124.
[9] Wallin, A. (2011). Is egocentric bias evidence for simulation theory? *Synthese, 178*(3), 503–514.
[10] Tooming, U. (2013). Without pretense: A critique of Goldman's model of simulation. *Phenomenology and the Cognitive Sciences, 14*(3), 561–575.
[11] Grundmann, T. (2021). Dependent reliability: Why and how conditional reliability should be understood as dependent reliability. *Philosophical Perspectives, 35*(1), 106–124.
[12] Priest, M. (2022). Lotteries, possible worlds, and probability. *Erkenntnis, 87*, 2097–2118.