A Report on “Indirect Reciprocity with Private, Noisy, and Incomplete Information” by Hilbe et al. (2018)

Reviewer 2

February 11, 2026

v2

isitcredible.com

Disclaimer

This report was generated by large language models, overseen by a human editor. It represents the honest opinion of The Catalogue of Errors Ltd, but its accuracy should be verified by a qualified expert. Comments can be made here. Any errors in the report will be corrected in future revisions.

> I am wiser than this person; for it is likely that neither of us knows anything fine and good, but he thinks he knows something when he does not know it, whereas I, just as I do not know, do not think I know, either. I seem, then, to be wiser than him in this small way, at least: that what I do not know, I do not think I know, either.
>
> Plato, *The Apology of Socrates*, 21d

> To err is human. All human knowledge is fallible and therefore uncertain. It follows that we must distinguish sharply between truth and certainty. That to err is human means not only that we must constantly struggle against error, but also that, even when we have taken the greatest care, we cannot be completely certain that we have not made a mistake.
>
> Karl Popper, 'Knowledge and the Shaping of Reality'

Overview

Citation: Hilbe, C., Schmid, L., Tkadlec, J., Chatterjee, K., and Nowak, M. A. (2018). Indirect Reciprocity with Private, Noisy, and Incomplete Information. *Proceedings of the National Academy of Sciences*, Vol. 115, No. 48, pp. 12241–12246.

Abstract Summary: This research explores indirect reciprocity when information about others' actions is private and noisy, contrasting with past research that assumed publicly available and synchronized information. The study finds that under these conditions, most of the eight crucial moral systems (leading-eight strategies) that maintain cooperation under public information cease to be stable, and those that do evolve are unable to sustain full cooperation.

Key Methodology: Evolutionary game theory, Markov chain simulations, and analytical recovery analysis are used to model reputation dynamics and strategy evolution in a well-mixed population.

Research Question: Which social norms can maintain stable cooperation in a society when information about other population members is private and noisy?

Editor's Note

Version 2 of this report has been written by an improved model of Reviewer 2.

Summary

Is It Credible?

The article by Hilbe et al. serves as a critical stress test for established theories of the evolution of cooperation. Previous research identified "leading-eight" social norms—strategies such as "Stern Judging"—that theoretically stabilize cooperation. However, these models relied on the assumption that reputation information is public and synchronized. Hilbe et al. challenge this by introducing a model where information is private, noisy, and incomplete. Their central claim is that under these more realistic conditions, most of these leading strategies fail to maintain stability, and those that do survive yield significantly reduced cooperation rates.

The credibility of the authors' primary negative result—the destabilization of strict moral norms—is high. By employing both evolutionary simulations and analytical modeling of reputation dynamics, the study convincingly demonstrates the fragility of strategies like Stern Judging (L6) and Judging (L8). The mechanism identified is logical and robust within the model's framework: when perception errors occur in a private information setting, disagreements about a third party's reputation emerge. Strict norms that mandate punishing "bad" actors cause these disagreements to proliferate, as one observer's "justified punishment" is viewed by another as "unjustified defection."

Crucially, the authors distinguish between "strict" and "generous" assessment rules. They show that generous strategies (like L1 and L3), where a cooperating donor is always perceived as good regardless of the recipient's standing, recover more quickly from errors than strict strategies like Stern Judging (L6), which judge cooperation against "bad" recipients as "bad." The analytical results regarding recovery times support this, showing that strict strategies struggle to recover from even a single disagreement (p. 12243). The simulations show that L1, L2, and L7 are the primary strategies capable of maintaining positive cooperation rates, but even these are not fully stable; rather than maintaining a high equilibrium, they often fall into dynamic cycles (rock-paper-scissors dynamics) involving unconditional cooperators and defectors (p. 12244). The finding that even the most resilient strategies see cooperation drop below 70% when error rates exceed 5% effectively quantifies the limitations of indirect reciprocity in noisy environments (p. 12244).

However, the study's conclusions regarding the solution to this problem require careful interpretation. The authors assert that their findings "highlight the importance of coordination and communication for the stability of indirect reciprocity" (p. 12245). It is important to note that this is a logical inference derived *ex negativo*, rather than a mechanism directly tested by the model. The model explicitly assumes that individuals are "completely independent when forming their beliefs" and are not "engaged in communication" (p. 12245). While the authors bolster this claim by citing external experimental evidence suggesting gossip aligns beliefs, the model itself does not empirically demonstrate that communication mechanisms (such as gossip or public tribunals) would successfully restore stability within these specific constraints (p. 12245). The necessity of communication is a plausible hypothesis generated by the failure of the independent-observer model, but the dynamics of that communication remain outside the scope of this specific analysis.

Additionally, the reliance on a "well-mixed population" assumption introduces a specific boundary condition to the results. The authors hypothesize that network-structured populations might amplify the problem of incomplete information because players can only observe immediate neighbors (p. 12245). This is a valid concern, yet it is also possible that the well-mixed assumption represents a severe test case in its own right, as global interactions allow errors to propagate across the entire population, whereas network clusters might contain the damage of reputation divergence. While this does not undermine the validity of the failure of strict norms in the presented context, it suggests that the universality of the collapse might depend heavily on the specific topology of social interactions—a variable not fully explored here. Nevertheless, the article succeeds in its primary goal: demonstrating that the "leading-eight" strategies are not the robust, universal solutions to cooperation they were previously thought to be when information is imperfect.

The Bottom Line

Hilbe et al. provide a highly credible refutation of the idea that strict social norms alone can sustain cooperation when reputation information is private and noisy. The analysis robustly demonstrates that strategies like "Stern Judging" collapse under perception errors due to the proliferation of disagreements, while even "generous" strategies often succumb to cyclic instability. However, the study's conclusion that communication is the necessary remedy is a logical inference rather than a tested result, as the model simulates only the absence of such coordination.

Potential Issues

The conclusion about communication is an inference not directly tested by the model: The article concludes that its findings "highlight the importance of coordination and communication for the stability of indirect reciprocity" (p. 12245). This conclusion is presented as an interpretation of the model's results, which demonstrate that cooperation is unstable when individuals form their beliefs independently under noisy conditions. The model is explicitly designed to explore this scenario by structurally excluding any form of communication, gossip, or consensus-building that could resolve disagreements about reputations. As the authors state, "The individuals in our model are completely independent when forming their beliefs" (p. 12245). While they reference external literature to support the utility of gossip, the study's purpose is to test the robustness of cooperative strategies *in the absence* of such synchronizing mechanisms. Therefore, the conclusion about the importance of communication is a logical inference about what is needed to restore stability, rather than a result that is or could be tested by the model's own methods.

The well-mixed population assumption may represent a less severe test case than structured populations: The study's findings, particularly the dramatic failure of strict norms like L6 ("stern judging"), are derived from a "well-mixed" population model where any two individuals are equally likely to interact. This combination of global interaction potential and isolated private information creates a scenario where perception errors can propagate widely. However, the authors themselves suggest that this assumption may not represent the worst-case scenario. They speculate that the negative effects of private information "might be even more pronounced when games take place on a network" because network structures can amplify the problem of incomplete information by limiting who can observe whom (p. 12245). The authors' argument implies that the well-mixed assumption, where every player has some chance of observing every other player's interactions, might be a more conservative test of these norms’ stability compared to a network where information is structurally siloed.

Future Research

Modeling explicit communication mechanisms: Future work should directly test the authors' inference by incorporating communication channels into the noisy information model. Research could introduce variables for "gossip" or "consensus building," where agents periodically synchronize their reputation scores with neighbors. This would determine if communication actually restores the stability of "leading-eight" strategies or if the noise in the communication channel itself introduces new instabilities.

Topological sensitivity analysis: To address the ambiguity regarding population structure, researchers should replicate the noisy/private information dynamics on various network topologies (e.g., small-world, scale-free, and lattice networks) rather than solely well-mixed populations. This would clarify whether local clustering acts as a firewall against the spread of reputation disagreements or, as the authors hypothesized, amplifies the issues of incomplete information by restricting observation.

Continuous vs. binary reputation dynamics: The current model relies on binary "good/bad" reputations, which may contribute to the brittleness of the strict strategies. Future research could implement continuous reputation scores (e.g., 0 to 1) or "forgiveness" thresholds. This would test whether a more nuanced, non-binary assessment system allows for greater resilience against the perception errors that destabilized the strategies in this study.

isitcredible.com