

Let Them Talk: A Multi-Level Analysis of Communication Between LLM Agents

Chaipat (Tata) Tirapongprasert
Columbia University

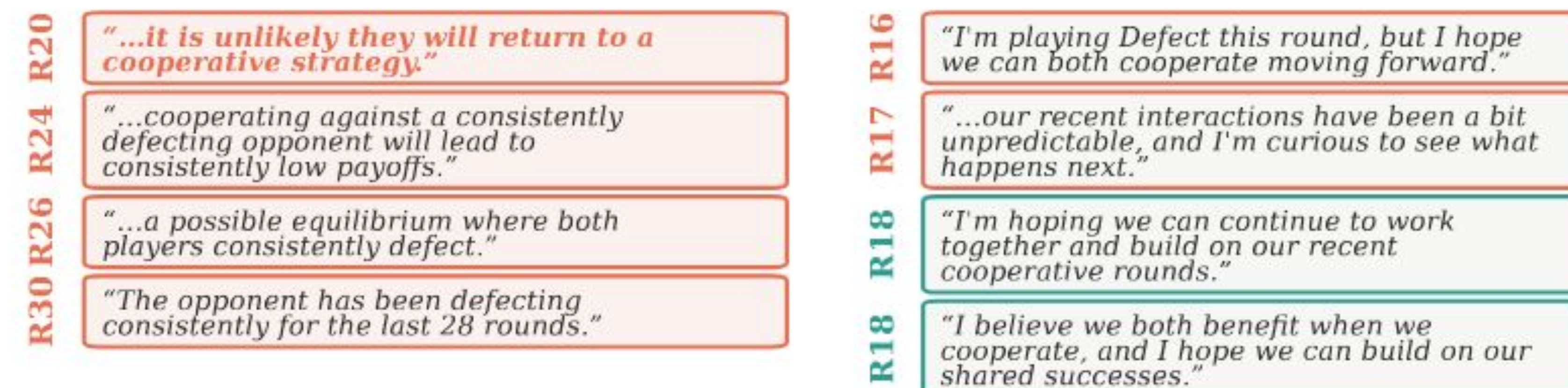
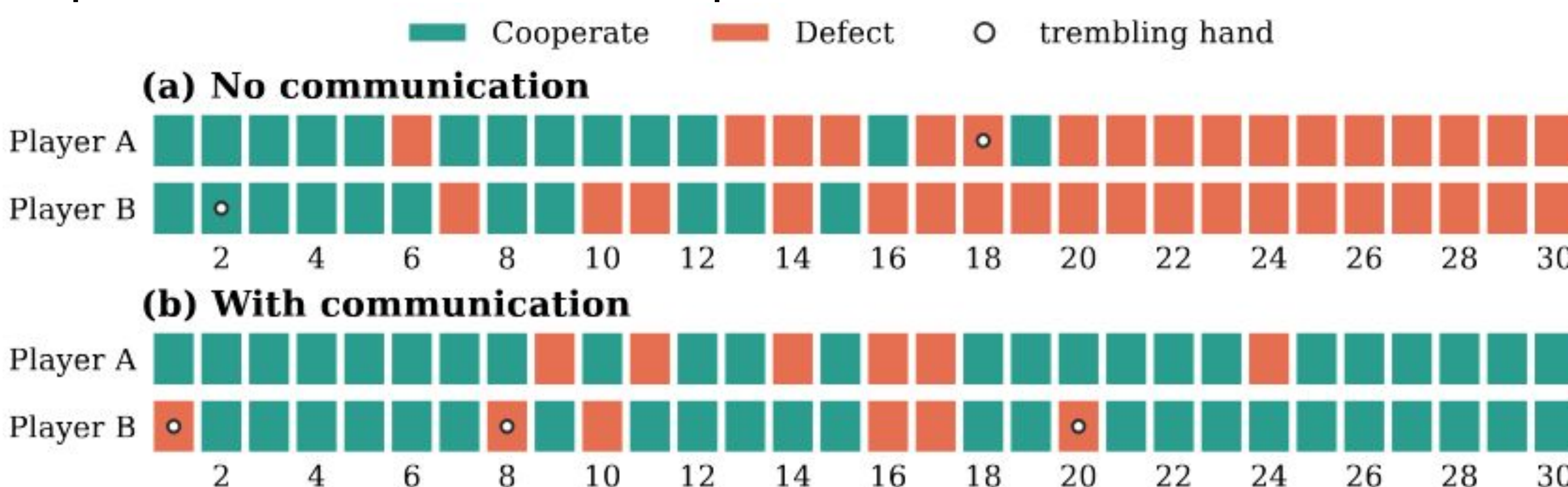
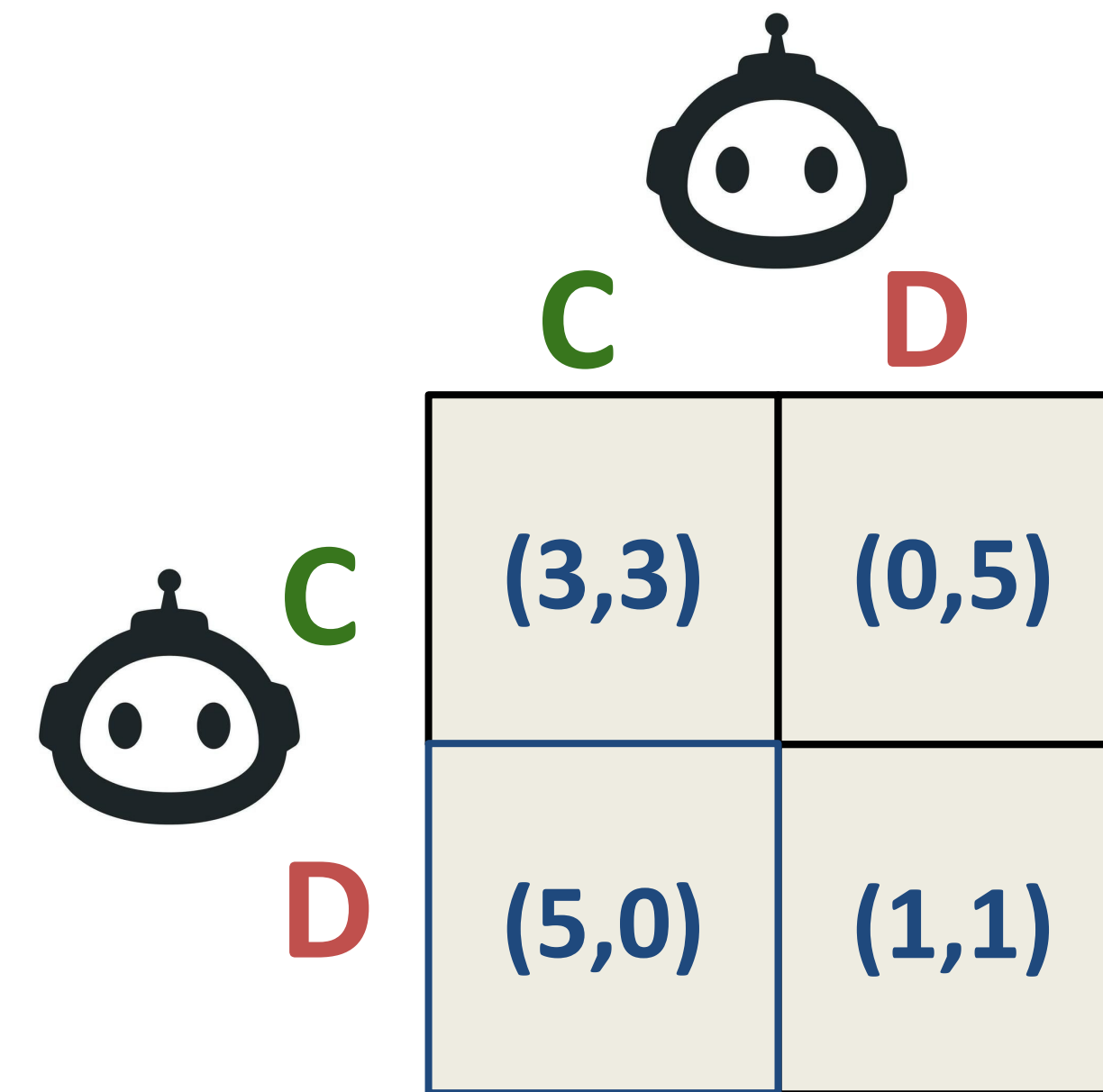
Rajarshi Das
MQube Cognition

Parijat Dube
Columbia University

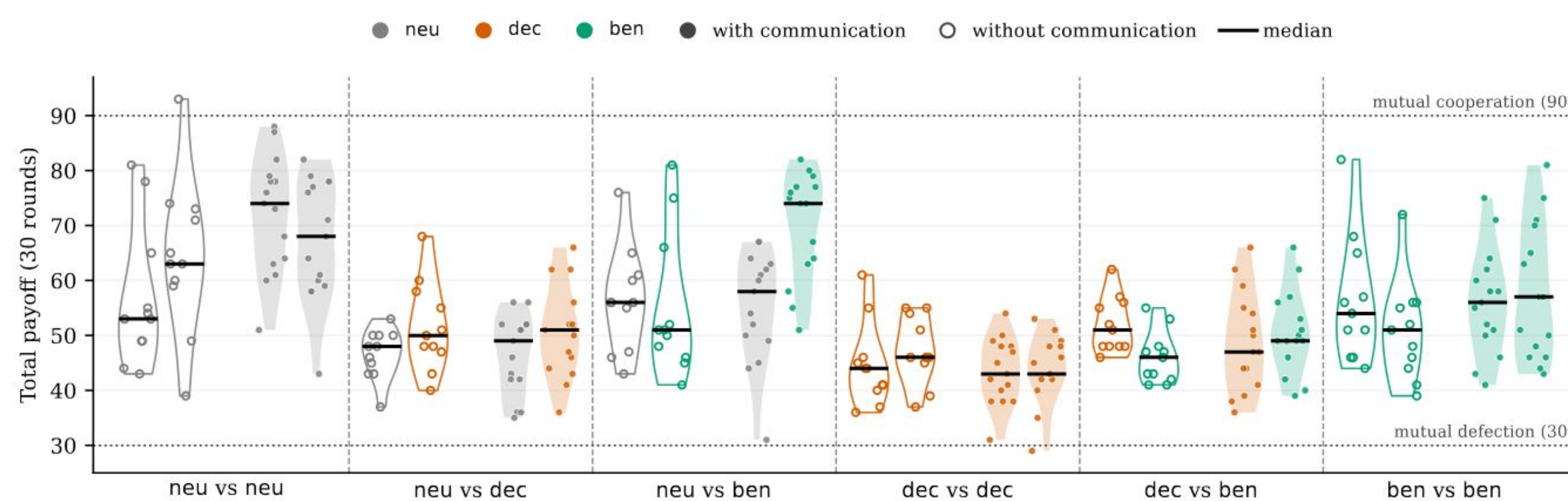
How communication affects decision-making

Problem: In the Iterated Prisoner's Dilemma, each rational agent wants to defect, despite both parties winding up worse off than if they had cooperated. Mutual cooperation relies on trust, which for both humans and LLMs is difficult to restore after the opponent's betrayal.

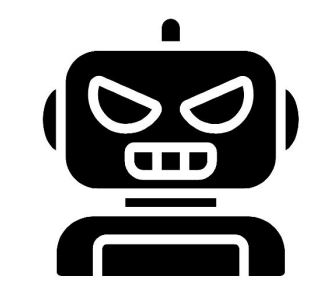
Claim: LLM agent-to-agent communication can improve coordination and promote trust in repeated interactions with incomplete information.



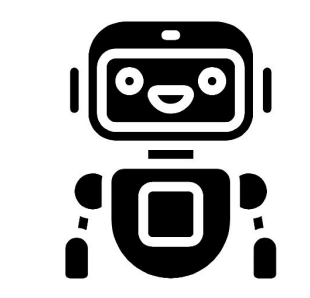
If an agent enters the **D** state, either on purpose or unintentionally, the opponent may respond to a stale, adversarial history, leading to mutual defection. **Communication acts as a synchronization channel.** It enables a well-intentioned agent to recover from this suboptimal state into **C** by signaling intent and repairing reputations following noisy defections.



Robustness to malicious actors

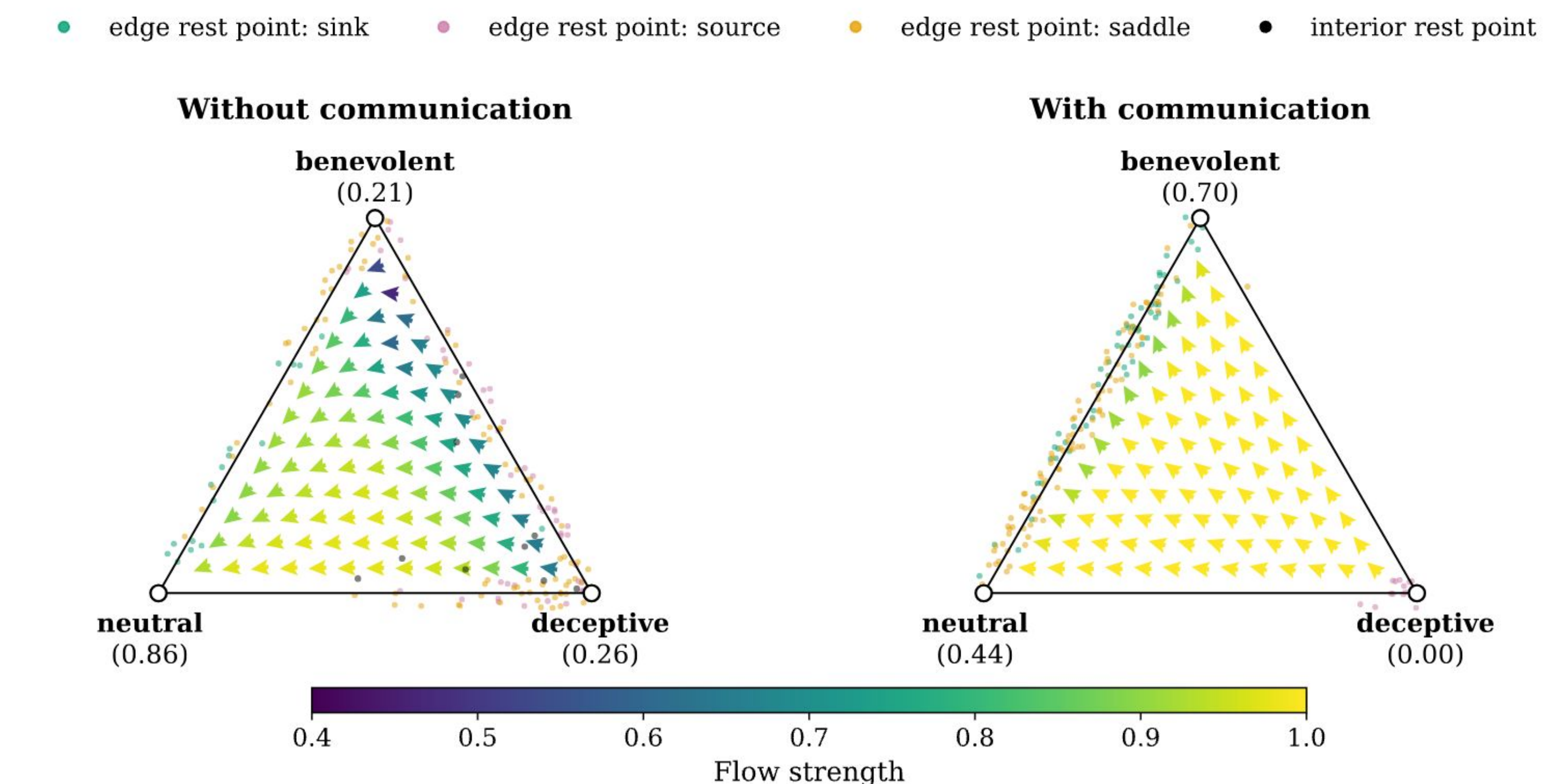


Dec: You are self-interested: you care more about your own payoff than about your opponent's. You may consider how to mislead or deceive your opponent, if you think it serves your interest.

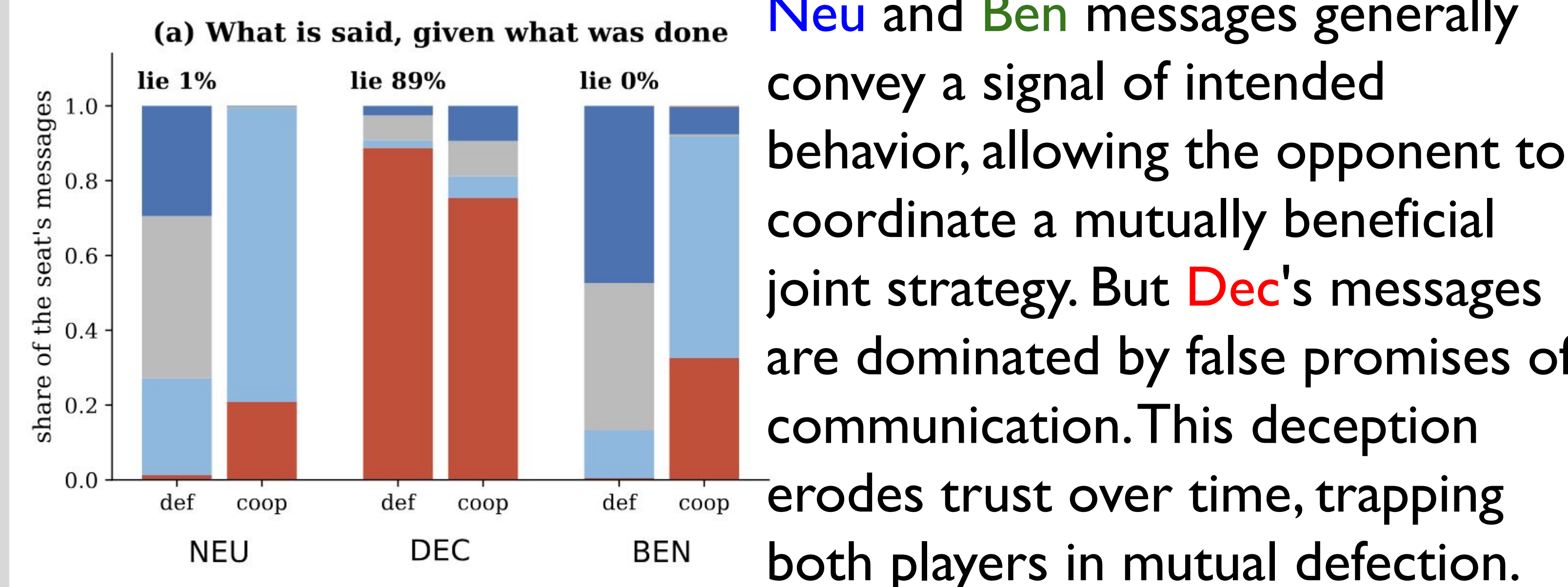


Ben: You are cooperative: you care about your opponent's payoff as much as your own, and you do not wish to deceive your opponent, but you will defend your score if they try to exploit it.

Result: Without communication, the replicator flow fragments into competing basins. Communication reshapes the payoff structure into a coherent flow toward neutral and benevolent cooperators.



Honest communication improves payoff.

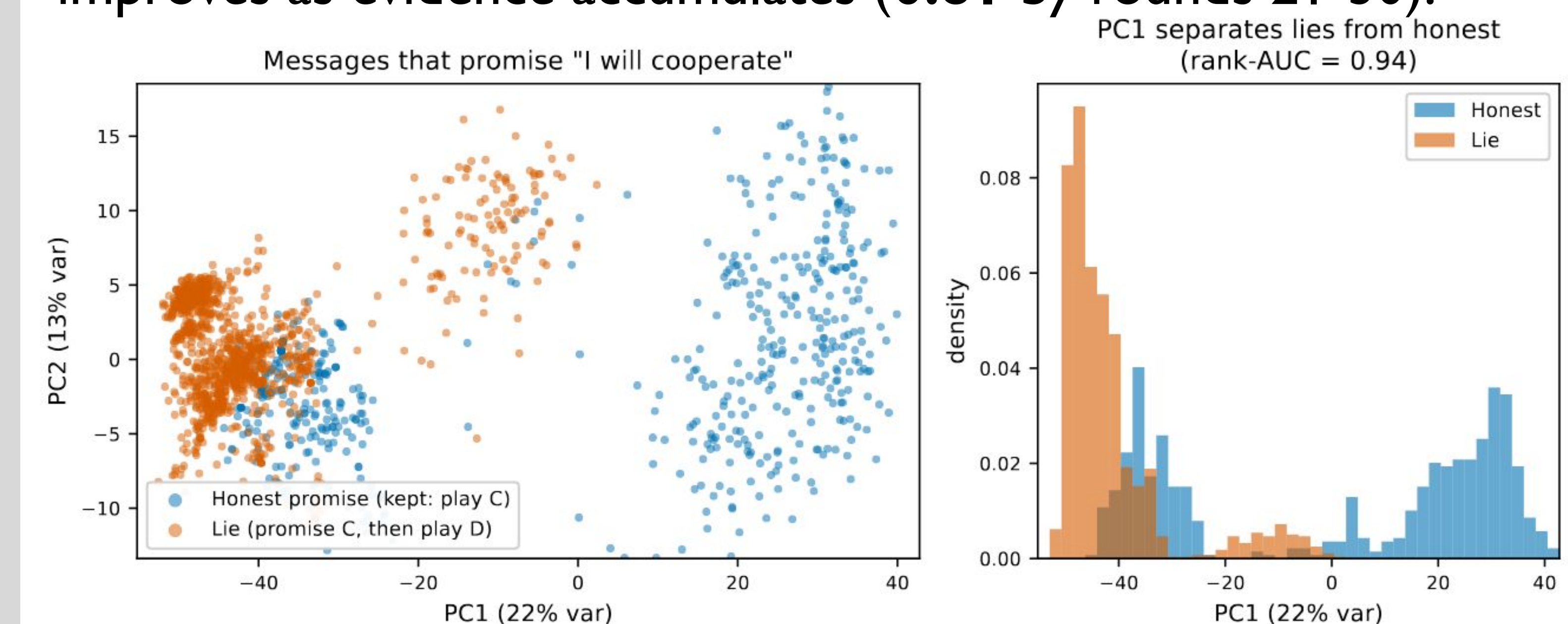


Communication makes honest agents more forgiving. A cooperative agent can only afford to forgive a defection if it can distinguish between an honest mistake and intentional exploitation. Since **Neu**'s and **Ben**'s messages provide useful information about their next actions, a partner can attribute the occasional defection to trembling noise and later re-extend cooperation.

LLMs form and update beliefs about the credibility of their opponents throughout the game.

To evaluate if the LLM agents store representations about the game and use them to guide the final action, we cache the residual stream at layer 32 and train logistic probes.

Our probes reveal that the opponent's type is decodable at the message token (**AUC 0.77** in rounds 1-10) and that its signal improves as evidence accumulates (**0.81** by rounds 21-30).



Deception is a mismatch between the message and the internal state. A true act of honest signaling aligns the agents' words with their internal representation: the probe detects defection in only 1% and 0% of rounds for **Neu** and **Ben**.

Conversely, even as **Dec** is crafting a comforting sentence to its opponent, PCA reveals that its own representation has already encoded the intended betrayal.

Caveat. The channel also makes the population more exploitable to **free riders** when the benefit of cooperation is non-excludable, like in the Public Good Game. This is because the benevolent majority still needs to keep the commons intact despite its refusal to contribute.

