Maciej Chudek, University of British Columbia, Canada
Joseph Henrich, University of British Columbia, Canada

How exploitation launched human cooperation: A model of negative indirect reciprocity

The evolution of cooperative human communities remains a central scientific puzzle. Most models emphasize positive reciprocity or coordinated punishment. These models target the latter stages of the evolution of human sociality as they assume the existence of (1) well-defined social roles (e.g., donors or punishers) or institutions (e.g., punishment pools or signal meaning), (2) sophisticated cognitive abilities for recognizing and responding to these socially defined roles and expectations, and (3) harmonious communities, where the benefits of mutual aid are not undermined by mutual exploitation.

Here we lay a foundation for these latter stages by describing how Negative Indirect Reciprocity (NIR) can suppress exploitation (e.g., stealing, rape, etc.) to (1) create harmonious communities, and (2) sustain supporting public goods contributions, while at the same time producing conditions that favour a sensitivity to shared social expectations, and the ability to recognize violations of these expectations. This negative context, where 'cooperating' means 'not exploiting someone', poses a distinct challenge since such cooperative acts are unobservable. While existing models depend on efficient helping (high benefits at low cost), NIR models stabilise harmonious communities when exploitation is inefficient (victims are seriously hurt by exploiters who benefit little, a large net loss). These inefficient circumstances are potentially more plausible earlier in human evolutionary history.

Our approach recasts the emergence of cooperative communities by showing how NIR can suppress opportunistic exploitation (e.g., of the weak, sick or injured) and then selectively revoking this safety net to enforce costly adherence to community-wide cooperative expectations. These models provide novel insights and suggest new challenges to the evolution of our species' distinct form of sociality.