LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

STRATEGIC MINDS: THE GAME THEORY OF COOPERATION, COORDINATION AND COLLABORATION

INDIRECT RECIPROCITY

YOU SCRATCH MY BACK, THEY'LL SCRATCH YOURS

Adrian Haret
a.haret@lmu.de

June 3, 2024

RICHARD D. ALEXANDER

Moral systems are systems of indirect reciprocity.

By *moral systems* I mean rewards and punishment [...] to control social acts that, respectively, help or hurt others.

Alexander, R. D. (1987). *The Biology of Moral Systems*. Aldine Transaction.

In our abstracted view of social interactions as a series of Prisoner’s Dilemmas, we can assume that reward is cooperation, while punishment is defection.

RICHARD D. ALEXANDER

Moral rules are established and maintained primarily by application of the concepts of right and wrong.

Alexander, R. D. (1987). *The Biology of Moral Systems*. Aldine Transaction.

That’s to say, agents have rules (i.e., strategies) for how to mete out cooperation and defection.

RICHARD D. ALEXANDER

The question is thus raised: what must be added to the conflicts of interest that characterize all life to create the conditions sufficient to produce systems involving ethical and moral questions?

Alexander, R. D. (1987). *The Biology of Moral Systems*. Aldine Transaction.

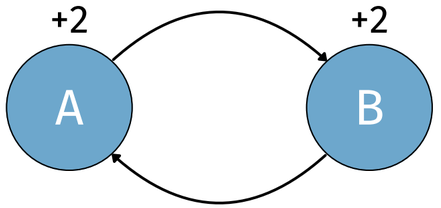
Translation: how can we model more complex strategies, based on indirect reciprocity?

But what *is* indirect reciprocity?

WILLIAM TRIVERS

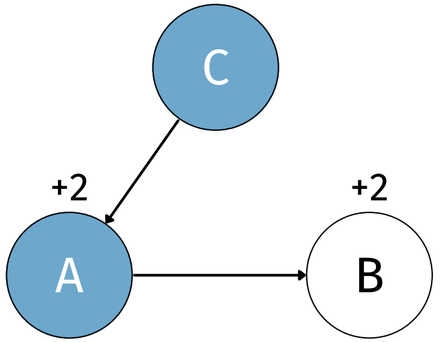
Direct reciprocity is I scratch your back, you scratch mine.

A confers a benefit to B, and B confers a benefit to A in return.



RICHARD D. ALEXANDER

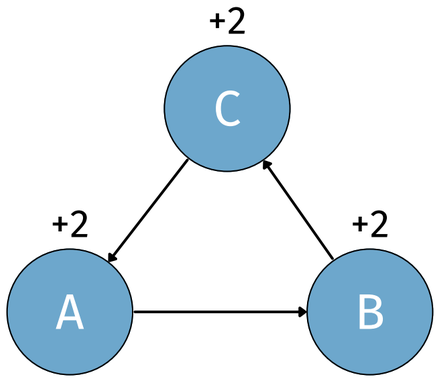
In indirect reciprocity, I scratch your back and someone else scratches mine.



Wait, wut? What’s the mechanism here?

RICHARD D. ALEXANDER

How about a kind of pay it forward mechanism:
A helps B, B helps C, C helps A.



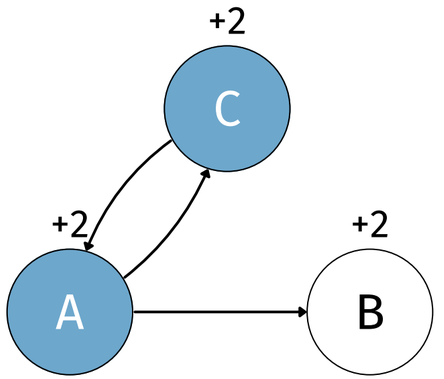
RICHARD BOYD

Well we tried to model this and it doesn't really work. :(

Boyd, R., & Richerson, P. J. (1989). The evolution of indirect reciprocity. *Social Networks*, 11(3), 213–236.

RICHARD D. ALEXANDER

Ok, how about this then.



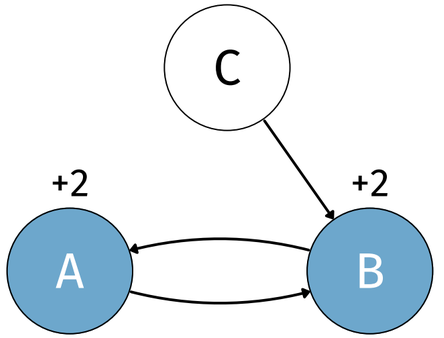
A helps B.

C, observing, later helps A.

A helps C.

RICHARD D. ALEXANDER

Or this, an example of altruism spreading.



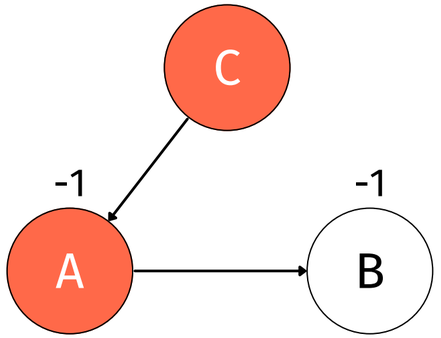
A helps B.

B helps A.

C, observing, helps B (expecting B will reciprocate).

RICHARD D. ALEXANDER

Or, an example of punishment spreading.



A hurts B.

C, observing, punishes A (expecting that if A goes unpunished, A will also hurt C).

RICHARD D. ALEXANDER

Rules for how and when to help/punish (i.e., systems of indirect reciprocity) are the basis for our moral systems!

Note that they require memory, consistency across time, the application of precedents, and persistent and widely communicated concepts of right and wrong.

DAVID HAIG

For direct reciprocity, you need a face; for indirect reciprocity, you need a name.

RICHARD D. ALEXANDER

Language and gossip come into play.

RICHARD D. ALEXANDER

Indirect reciprocity involves reputation and status, and results in everyone in a social group continually being assessed and reassessed by interactants, past and potential, on the basis of their interactions with others.

But does this work from an evolutionary (or game theory) perspective?

Like, why would C do any punishing? What's in it for them?

RICHARD BOYD

Yeah we’re skeptical.

MARTIN NOWAK

Well, maybe it can work...