Cooperation and Learning in Unfamiliar Situations

William H.B. McAuliffe[1], Maxwell N. Burton-Chellew[2], and Michael E. McCullough[1]

[1] Department of Psychology, University of Miami

[2] Department of Ecology and Evolution, University of Lausanne

Abstract

Human social life is rife with uncertainty. In any given encounter, one can wonder whether cooperation will generate future benefits. Many people appear to resolve this dilemma by cooperating initially, plausibly because (1) encounters in everyday life often have future consequences and (b) alienating oneself from long-term social partners, at least in our evolutionary history, often outweighs the short-term benefits of acting selfishly. However, since cooperating with others does not always advance self-interest, people might also learn to withhold cooperation in certain situations. Here, we review evidence for two ideas: that people (1) initially cooperate or not depending on the habits and incentives that they have developed during their daily lives; and (2) also learn through experience to adjust their cooperation towards the incentives of unfamiliar situations. We evaluate these claims in comparison with the widespread view that anonymously helping strangers in laboratory settings is motivated by altruistic desires. We conclude that the evidence is more consistent with the idea that people stop cooperating in unfamiliar situations because they learn that it does not help themselves, either financially or through social approval.

Keywords: cooperation, prediction error, economic games, trust, habit

People often help people they do not know or will never see again, donating blood to unknown beneficiaries, tipping taxi drivers in foreign cities, and posting reviews of hotels and restaurants they attend at conferences. Such behaviors are even present in tightly controlled, anonymous, laboratory experiments that strictly prevent any future interactions. What explains humans' propensity to expend resources to benefit others (to "cooperate") even when they do not know them and are unlikely to meet them again? One promising explanation is that our modern-day decency reflects the importance of maintaining relationships in our evolutionary past (Krasnow & Delton, 2016; Raihani & Bshary, 2015). This theory suggests that an instinct to cooperate still operates in today's interactions with strangers because it was beneficial to avoid alienating potential long-term partners in ancestral environments where people lived in small, close-knit communities.

However, cooperation is not always in one's long-term self-interest. Sometimes social partners prove untrustworthy, sometimes the gain from selfishness outweighs the cost of angering a social partner, and sometimes it really is unlikely that two people will ever meet again. Therefore, natural selection may have favored cognitive mechanisms that switch cooperative tendencies on or off accordingly. Because the incentives of a situation can be opaque, it is also likely that natural selection will have favored an ability to learn from experience when to modify cooperative tendencies. Here we review an emerging literature that suggests people can override default tendencies to cooperate, especially if they have opportunities to habituate to situations in which cooperation does not advance self-interest.

The Development of Social Prudence

A fear of being excluded will not curb selfishness among those who do not yet know that they will receive sanctions for behaving uncooperatively. Fortunately, human development

provides ample opportunities to learn the incentive structures of recurrent social situations. Young children receive commands from their parents to alter their behavior every eight minutes or so (Hoffman, 2000). Community members hear stories passed down from one generation to another that communicate the virtues of cooperation and how it is enforced (Smith et al., 2017). Even generally law-abiding citizens occasionally commit misdeeds (Gabor, 1994); and the negative consequences (or lack thereof) that they experience, or observe others experiencing, affects their likelihood of behaving badly in the future (Gächter & Schulz, 2016). Indeed, people on average become more cooperative with age, perhaps because experience teaches them that cheating is a losing strategy in the long run (Matsumoto, Yamagishi, Li, & Kiyonari, 2016).

Domain-General Process, Domain-Specific Knowledge

How do people incorporate the wisdom they have accrued over development into cooperation decisions? Following other recent researchers, we propose the concept of a *prediction error* to explain how people learn whether to cooperate in particular situations (FeldmanHall & Dunsmoor, 2018). A prediction error represents the discrepancy between the expected outcome of a decision (e.g., “I believe that cooperating in this situation will be good for my reputation in the long-term”) and the actual outcome (e.g., “It turns out that cooperating had no long-term effect on my reputation”). By incorporating past prediction errors in decision-making, people can choose whether to cooperate based on the presence or absence of situational factors that were correlated with desired outcomes in past experiences.

Prediction error learning is involved in revising many types of beliefs in light of new evidence, not just beliefs about whether cooperation is prudent. However, learning whether to cooperate is a domain-specific task inasmuch as people bring knowledge to bear that it is particularly useful for managing social relations. We theorize that people represent such

knowledge in their *initial* belief about whether cooperation advances long-term self-interest in any particular situation. For instance, people prefer interacting with in-group members from whom they anticipate a level of partiality that they do not expect from out-group members (Foddy, Platow, & Yamagishi, 2009). As a result, prediction errors will be largest when out-group members prove benevolent and when in-group members prove uncooperative. Moreover, the type of relationship that one has with an in-group member determines what constitutes acceptable behavior. For example, although reciprocation of specific deeds is expected in formal arrangements such as carpools, failing to pay back any one favor will typically engender less opprobrium in a familial relationship (Fiske, 1992).

The effect of prediction errors on subsequent cooperation decisions depends also on the *certainty* of initial beliefs. For example, Siegel, Mathys, Rutledge, and Crockett (2018) argued that it is not adaptive to reject potential social partners based on a few minor infractions because even cooperative people occasionally behave poorly. Instead, people should reserve judgment towards possibly remorseful transgressors. Consistent with such reasoning, the authors found that people make weaker inferences about moral character when observing uncooperative behavior than when observing cooperative behavior. This asymmetry caused participants to more quickly change their judgment of previously selfish individuals, who had begun cooperating, than of previously cooperative individuals who had begun behaving selfishly. These examples demonstrate that although the flexibility of cooperation decisions depends on a domain-general learning process, such learning occurs in the context of beliefs and desires that are specific to cooperative interactions.

Learning to (Not) Cooperate

One type of situation in which prediction errors are likely to influence future behavior are those in which it is never in one's self-interest to cooperate. For example, many *economic game* experiments require laboratory participants to decide whether to share windfalls of money with anonymous strangers in one-off interactions. Economic games present situations that differ greatly from those of everyday life, where people typically decide whether to share resources with others they already know or might interact with again. Thus, although cooperation in economic games could reflect a desire to benefit anonymous strangers, it could also arise from participants who have yet to register the mismatch between the situations they are familiar with and the unfamiliar economic game.

To illustrate, consider the debate over the existence of so-called conditional cooperators, who choose to cooperate based on whether they believe others will cooperate too. Researchers have inferred the existence of conditional cooperators from experiments in which people play a public goods game, where members of a group can contribute money to a common resource that benefits everyone equally. In the game, each dollar contributed benefits the group because the experimenter multiplies all contributions by a constant ($M>1$) before sharing them out equally. Thus, each dollar contributed returns M dollars to the group, and M/N dollars to each member. However, M is typically set to less than the number of group members ($N > M > 1$), so each dollar contributed is personally costly ($N/M < 1$). The canonical result from games with several rounds of decision-making is that conditional cooperators contribute much of their endowment in early rounds, but contribute little by the end (for a review, see Chaudhuri, 2011).

Many researchers posit that conditional cooperators want to promote the public good, but eventually stop contributing out of disgust with "free-riders" who seldom contribute (Fehr &

Schurtenberger, 2018). In essence, the conditional cooperation hypothesis posits that individuals perfectly understand the game, but must learn whether their counterparts share their goals. This hypothesis is challenged, however, by the fact that many conditional cooperators later report that their contribution decisions were based on a desire to maximize their own income (Burton-Chellew, El Mouden, & West, 2016). Furthermore, the same decline in cooperation occurs when participants play with computers rather than humans (Burton-Chellew & West, 2013; Houser & Kurzban, 2002). Thus, declining contributions in these cases cannot reflect a regard for others that is eroded by exasperation toward non-cooperators. Instead, declining contributions likely reflect a mistaken belief that contributions increase personal income when other participants also contribute, a confusion that is corrected by experience.

Further support for the confusion hypothesis comes from a study where participants played a public goods game with humans but could only observe the contributions of people in other groups, not in their own (Burton-Chellew, El Mouden, & West, 2017a). In these cases, participants mimicked the behaviors of successful players in other groups by reducing their contributions in their own groups. These findings suggest that people were using social information to learn how to improve their payoffs, although it is also possible that they were making inferences about whether their own group members would cooperate based on whether members of other groups cooperate. The experimenters confirmed the payoff-learning hypothesis in a subsequent public goods game with computers instead of humans: Participants who had observed successful players contributed less toward the computers than those who had not observed successful human players, indicating that the former group had learned how to maximize their income better than had the latter group.

Habituation to Everyday Incentives

Participants in economic games probably believe initially that cooperation promotes self-interest because they are importing the behaviors that have been rewarded in their everyday lives to novel situations. For example, participants from societies where interactions among strangers are effectively regulated by law cooperate more in economic games (Stagnaro, Arechar, & Rand, 2017). If people really do impose these mental models upon economic games, then the same economic game may evoke different working models from real-life for people from different backgrounds. An experimental demonstration of this phenomenon comes from Study 2 of Stagnaro et al. (2017), in which participants began by playing several rounds of a public goods game. In some conditions, stingy contributions were punished, whereas in the control condition participants could behave selfishly with impunity. Participants who had been "enculturated" in the punishment version of the public goods game were later more cooperative in a second economic game that did not feature punishment.

If initial behavior in laboratory studies is shaped by a spillover from everyday life, then behavior at the end of experiments—that is, after a period of adjustment to the local incentives—should be more reflective of participants' underlying goals (Binmore, 1999). For example, cooperation typically declines in repeated public goods games. A study using samples from 16 different societies showed that this phenomenon is cross-culturally valid, consistent with a pancultural desire to learn how to increase one's own income (Herrmann, Thöni, & Gächter, 2008). Furthermore, when participants return to the lab after having previously experienced a public goods game, they tend to cooperate less the second time around (Conte, Levati, & Montinari, 2019).

In a longitudinal demonstration of spillover, McAuliffe, Forster, Pedersen, and McCullough (2018) had participants play cooperation games with anonymous strangers and privately donate to charity on two separate occasions. People contributed about 20% less to charity and strangers on the second occasion, except in the one game in which cooperation *can* yield a personal profit, even when played only once with a stranger. The researchers argued that participants acted on cooperative habits from everyday life during the first session, but by the time they arrived at the second session had learned that nobody would thank them for behaving fairly or scold them for behaving selfishly. Consistent with this hypothesis, decisions at the first but not the second session were positively associated with self- and peer-reports of cooperative traits, which do reflect how people behave in everyday social interactions (McAuliffe, Forster, Pedersen, & McCullough, 2019).

Individual Differences

If registering prediction errors helps to align behaviors with incentives, then a greater *ability* to learn should increase the speed with which people make cooperation decisions that are congruent with their true preferences. Indeed, Burton-Chellew et al. (2016) found that free-riders were the only subset of participants who reliably understood that cooperation does not maximize personal income in the public goods game. Another study revealed that Japanese adults who never shared money across multiple one-shot economic games scored higher on an intelligence test and self-reported greater skill in understanding social situations (Yamagishi, Li, Takagishi, Matsumoto, & Kiyonari, 2014). Barreda-Tarrazona, Jaramillo-Gutiérrez, Pavan, and Sabater-Grande (2017) reported that reasoning ability is associated with less cooperation in one-shot prisoner's dilemma games, and with *more* cooperation in repeated prisoner's dilemmas with the same partner (see also Jones, 2008). Both patterns are optimal: In the one-shot game there is no

self-interested reason to cooperate, whereas in the repeated game inducing a long-term partner to cooperate in earlier rounds is essential to securing the benefits of cooperation in later rounds. Similarly, Burton-Chellew, El Mouden, and West (2017b) showed that participants who understood that it was pointless to contribute in a public goods game with computer groupmates were the only ones to strategically increase their initial contributions in a repeated game with humans when their behavior was potentially observable to their groupmates.

Although we have criticized a readiness to interpret cooperation in unfamiliar situations as evidence of unselfishness, we hasten to note that learning does not inevitably facilitate selfishness. Instead, understanding a situation's reward structure should increase alignment between people's behavior and their goals, whatever those goals happen to be. For example, Lockwood, Apps, Valton, Viding, and Roiser (2016) had participants learn over several trials which of two choice options was associated with a monetary reward for either the self or another person. Overall, participants minimized prediction errors more quickly when the rewards went to the self. However, the discrepancy in speed between learning on behalf of the self and another person was smaller for people higher in trait empathy, suggesting that empathic individuals are motivated to learn how to help others.

Conclusion

Experimental economists have long emphasized the role of learning in social decision-making (e.g., Binmore, 1999). However, cooperation researchers have only recently considered how peoples' past social interactions shape their expectations in novel social situations. An important lesson from the research reviewed here is that people's behavior in any single situation is not necessarily a direct read-out of how selfish or altruistic they are, especially if the situation's incentives differ from what they normally encounter in everyday life.

References

Barreda-Tarrazona, I., Jaramillo-Gutiérrez, A., Pavan, M., & Sabater-Grande, G. (2017). Individual characteristics vs. experience: An experimental study on cooperation in prisoner's dilemma. *Frontiers in Psychology*, *8*:596. doi: 10.3389/fpsyg.2017.00596

Binmore, K. (1999). Why experiment in economics? *The Economic Journal*, *109*(453), 16-24.

Burton-Chellew, M. N., El Mouden, C., & West, S. A. (2016). Conditional cooperation and confusion in public-goods experiments. *Proceedings of the National Academy of Sciences*, *113*(5), 1291-1296.

Burton-Chellew, M. N., El Mouden, C., & West, S. A. (2017a). Social learning and the demise of costly cooperation in humans. *Proceedings of the Royal Society of London B: Biological Sciences*, *284*(1853), 20170067.

Burton-Chellew, M. N., El Mouden, C., & West, S. A. (2017b). Evidence for strategic cooperation in humans. *Proceedings of the Royal Society B: Biological Sciences*, *284*(1856), 20170689.

Burton-Chellew, M. N., & West, S. A. (2013). Prosocial preferences do not explain human cooperation in public-goods games. *Proceedings of the National Academy of Sciences*, *110*(1), 216-221.

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics*, *14*(1), 47-83.

Conte, A., Levati, M. V., & Montinari, N. (2019). Experience in public goods experiments. *Theory and Decision*, *86*(1), 65-93.

Fehr, E., & Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour*, *2*(7), 458.

FeldmanHall, O., & Dunsmoor, J. E. (2018). Viewing adaptive social choice through the lens of associative learning. *Perspectives on Psychological Science*, 1745691618792261.

Fiske, A. P. (1992). The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review*, *99*(4), 689-723.

Foddy, M., Platow, M. J., & Yamagishi, T. (2009). Group-based trust in strangers: The role of stereotypes and expectations. *Psychological Science*, *20*(4), 419-422.

Gabor, T. (1994). *" Everybody Does It!": Crime by the Public*. University of Toronto Press.

Gächter, S., & Schulz, J. F. (2016). Intrinsic honesty and the prevalence of rule violations across societies. *Nature*, *531*(7595), 496-499.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, *319*(5868), 1362-1367.

Hoffman, M. L. (2000). *Empathy and moral development: Implications for caring and justice*. Cambridge University Press.

Houser, D. & Kurzban, R. (2002). Revisiting kindness and confusion in public goods experiments. *American Economic Review*, *92*, 1062-1069.

Jones, G. (2008). Are smarter groups more cooperative? Evidence from prisoner's dilemma experiments, 1959–2003. *Journal of Economic Behavior & Organization*, *68*(3-4), 489-497.

Krasnow, M. M., & Delton, A. W. (2016). Are humans too generous and too punitive? Using psychological principles to further debates about human social evolution. *Frontiers in Psychology*, *7*, 799.

Lockwood, P. L., Apps, M. A., Valton, V., Viding, E., & Roiser, J. P. (2016).

Neurocomputational mechanisms of prosocial learning and links to empathy. *Proceedings of the National Academy of Sciences*, *113*(35), 9763-9768.

Matsumoto, Y., Yamagishi, T., Li, Y., & Kiyonari, T. (2016). Prosocial behavior increases with age across five economic games. *PloS one*, *11*(7), e0158671.

McAuliffe, W.H.B., Forster, D.E., Pedersen, E.J., & McCullough, M.E. (2018). Experience with anonymous interactions reduces intuitive cooperation. *Nature Human Behaviour*, *2*, 909-914.

McAuliffe, W.H.B., Forster, D.E., Pedersen, E.J., & McCullough, M.E. (2019). Does cooperation in the laboratory reflect the operation of a broad trait? *European Journal of Personality*, *33*, 89-103.

Raihani, N. J., & Bshary, R. (2015). Why humans might help strangers. *Frontiers in behavioral neuroscience*, *9*, 39.

Siegel, J. Z., Mathys, C., Rutledge, R. B., & Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature Human Behaviour*, *2*(10), 750-756.

Smith, D., Schlaepfer, P., Major, K., Dyble, M., Page, A. E., Thompson, J., ... & Ngales, M. (2017). Cooperation and the evolution of hunter-gatherer storytelling. *Nature Communications*, *8*(1), 1853.

Stagnaro, M. N., Arechar, A. A., & Rand, D. G. (2017). From good institutions to generous citizens: Top-down incentives to cooperate promote subsequent prosociality but not norm enforcement. *Cognition*, *167*, 212-254.

Yamagishi, T., Li, Y., Takagishi, H., Matsumoto, Y., & Kiyonari, T. (2014). In search of Homo economicus. *Psychological Science*, *25*(9), 1699-1711.

Recommended Reading

Raihani & Bshary, 2015. (See references.) An overview of competing evolutionary explanations for why humans help others whom they are unlikely to interact with again.

Binmore, K. (1999). (See references.) An early review of experimental economics research suggesting that participants' true economic preferences do not emerge until they have had experience with the laboratory task at hand.

Fehr, E., & Schurtenberger, I. (2018). (See references.) A recent articulation of the view that conditional cooperation reflects unselfish motives (as opposed to the view advanced here, which is that conditional cooperation is an artefact of learning how to maximize one's own income).

FeldmanHall, O., & Dunsmoor, J. E. (2018). (See references.) A review of operant and Pavlovian learning processes that affect cooperation decisions.

Yamagishi, T., Hashimoto, H., & Schug, J. (2008). Preferences versus strategies as explanations for culture-specific behavior. *Psychological Science*, *19*(6), 579-584. An early demonstration that social behavior in study situations may reflect habituation to the incentive structure of everyday life rather than the incentives of the study situation.