Human Cooperation and the Complexity of Moral Codes

Fernando P. Santos[1,2], Jorge M. Pacheco[3,2], and Francisco C. Santos[1,2]

[1]GAIPS/INESC-ID and Instituto Superior Técnico, Universidade de Lisboa, Portugal
[2]ATP-group, Portugal [3]CBMA and Universidade do Minho, Portugal

Explaining the emergence of cooperation remains a fundamental challenge amongst many areas of science. When paying a cost (*e.g.*, money, time, energy) is required to help others, cooperation often constitutes a social dilemma: society benefits if everyone cooperates however, given the cost involved, individuals are tempted to defect. Theoretical and experimental works have shown that reputations may solve this cooperation conundrum. These features are often framed in the context of indirect reciprocity (IR), which constitutes the most elaborate, cognitively demanding, and specifically human mechanism of cooperation discovered so far. By helping someone, individuals may increase their reputation, which can change the predisposition of others to help them in the future. The reputation of an individual depends, in turn, on the social norms that establish what characterises a good or bad action, providing a basis for morality.

In this context, Ohtsuki and Iwasa analysed the stability of cooperative strategies under all possible 3rd order social norms (detailed next) and, interestingly, only eight distinct norms were found to stabilise high levels of cooperation (Ohtsuki and Iwasa, 2004, 2006). In reality, however, individuals may have access to past reputations of others, which opens space for more complex – and meticulous – social norms and strategies, with non-trivial effects on the evolution of cooperation under IR. In fact, norms based on indirect reciprocity can be sufficiently complex to challenge an individual's cognitive ability to follow moral rules. A simple social norm may say that cooperation always leads to a good reputation. A more complex social norm may postulate that cooperation with individuals that are good in the present but were bad in the past is right, whereas cooperation with whoever is bad now and was also bad in the past is wrong. This begs the question of how simple can a social norm be while still promoting cooperation in a society, when individuals have access to the previous reputations of their peers.

In **Santos et al. (2018b)** we provide an answer to this question, while proposing a new framework to analyse strategy and norm complexity of potential relevance to a broad range of decision-making problems. Our results are able to identify a key pattern of social norms, which leads to maximal cooperation at minimal complexity, more so if we consider a complexity cost in the decision process. This unique combination suggests that simple moral principles can excel in eliciting cooperation even in complex environments.

These results were obtained through large-scale agent-based simulations. We consider a binary world where reputations can be either Good (G) or Bad (B), leading to a large set of possible social norms, employed in the classification of decisions taken in an instance of a donation game: A donor may either Cooperate (C), helping a recipient at a cost c to herself/himself while conferring a benefit b to the recipient (with $b > c > 0$), or Defect (D, no help provided), in which case no one incurs any costs or distributes any benefits. Everyone in the population employs the same social norm to assign a public reputation to an individual. This reputation is attributed and is disseminated by a bystander who witnesses a pairwise interaction. In this context, norms that only consider the action of the donor are said to be 1st order norms. If, besides the donor action, the actual reputation of the recipient also matters, one obtains a 2nd order norm (Santos et al., 2016). A 3rd order norm further includes the actual reputation of the donor (Ohtsuki and Iwasa, 2004, 2006). To fully address the complexity of social norms, we propose a new space of 4th order norms, which further incorporates the previous reputation of the recipient. Under this setting, the strategy of each individual constitutes a policy that dictates cooperation/defection when interacting in different contexts, being represented by a tuple $p = (p_0, p_1, ..., p_7)$ in which $p_i \in \{0, 1\}$ denotes the action of the donor (C or D) for each of the possible combination of reputations: past reputation of recipient (G or B), actual reputation of donor and atual reputation of recipient. Likewise, a norm is given by a tuple $d = (d_0, d_1, ..., d_{15})$, in which $d_i \in \{0, 1\}$ denotes the new reputation assigned to the donor for each of the possible combination of action, past reputation of recipient, actual reputation of donor and atual reputation of recipient). This way, there are 2^{16} 4th order norms and 2^8 different strategies.

Equipped with these tools, we investigate which norms promote cooperation. We perform computer simulations where individuals in a population, each starting with a random strategy, play the donation game with their peers and eventually change strategies via social-learning. Strategies with higher fitness are more frequently adopted. The simulations return the cooperation index (η), i.e., the average fraction of donations observed in a population evolving under a given norm (Santos et al., 2016).

The results show that only a very small fraction of social norms (approximately 0.2% out of the 2^{16} norms analysed) are able to sustain levels of cooperation above 90%. Among those, we find that only part of the moral principles that leads to cooperation in the space of 3rd order norms remain equally efficient within a larger 4th order space. Among those norms that promote maximal levels ($> 90\%$) of cooperation, some are more complex than others. To quantify their complexity we look at social norms as Boolean functions. The complexity or of a norm (κ) is given by the length of the shortest logically equivalent Boolean formula (here in disjunctive normal form, DNF). A similar quantity was used in the past to describe Human's subjective difficulty of learning a concept (Feldman (2000)). Clearly, norms of the same order may entail different cognitive complexities. Moreover, similar to norms, strategies also exhibit an intrinsic complexity (κ_s) that may influence their adoption.

We found that some cooperative norms translate into extremely simple moral judgements and even overlook the past reputations of individuals. A simple 2nd order norm (known as *Stern-Judging*) has been found as one that maximizes cooperation in the artificial societies simulated, while minimising its cognitive complexity (see Fig. 1). Such a moral code stipulates that "*only whoever cooperates with good and defects with bad should have a good reputation*".

The success of stern-judging suggests that cooperation under IR may only require simple information processing mechanisms and norms easy to internalise. This is particularly relevant as IR relies on indirect (and erroneous) information about their peers. Showing that IR can sustain cooperation with simple rules and reduced information supports its significance in real systems. Interestingly, the fingerprint of stern-judging can be also found in recent developmental psychology research (Hamlin et. al. (2011)) showing that infants, since early ages, have a preference not only for characters who helped others, but also for characters who harmed those who hindered others. This suggests that the moral principles one resorts to at early developmental stages can suffice to ensure pro-social behaviours in societies.

The results are qualitatively insensitive to the ratio b/c, population size, errors in assessment or assignments made by individuals and different schemes of random exploration of strategies. Furthermore, one of the fundamental ingredients of indirect reciprocity is that individuals report their interactions. This naturally involves time and effort. When the process of information sharing is costly we show that stern-judging remains highly efficient, particularly when some sort of anticipation is at place (Santos et al., 2018a).

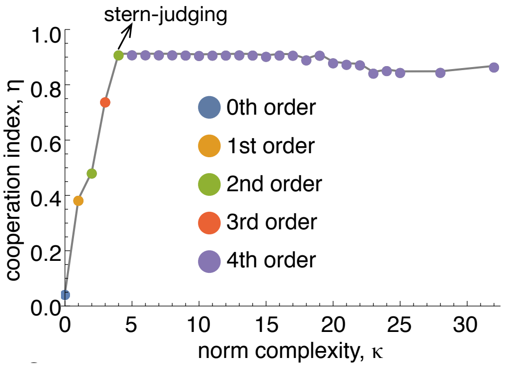

Figure 1: Maximal cooperation (η) for each complexity (κ) level. *Stern-judging* leads to $\eta > 0.9$ at low $\kappa = 4$.

The model developed provides a new perspective on IR and a new conceptual framework to investigate the complexity of social norms. Furthermore, this work may provide clues on policymaking and the design of reputation systems, pervasive in nowadays web platforms and systems supporting sharing economies. Finally, in the realm of societies comprising humans and artificial agents, these results may help to identify the core values to implement in autonomous agents to maintain and foster pro-social behaviours in such hybrid societies (Paiva et al. (2018)).

Acknowledgements. This work was supported by FCT Portugal (Grants: SFRH/BD/94736/2013, PTDC/EEI-SII/5081/2014, PTDC/MAT/STA/3358/2014, UID/BIA/04050/2013 and UID/CEC/50021/2013).

References

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, 407(6804):630–633.

Hamlin et. al. (2011). How infants and toddlers react to antisocial others. *PNAS*, 108:19931–19936.

Ohtsuki, H. and Iwasa, Y. (2004). How should we define goodness? - reputation dynamics in indirect reciprocity. *J Theor Biol*, 231(1):107–120.

Ohtsuki, H. and Iwasa, Y. (2006). The leading eight: social norms that can maintain cooperation by indirect reciprocity. *J Theor Biol*, 239(4):435–444.

Paiva, A., Santos, F. P., and Santos, F. C. (2018). Engineering prosociality with autonomous agents. In *AAAI'2018*.

Santos, F. P., Pacheco, J. M., and Santos, F. C. (2016). Evolution of cooperation under indirect reciprocity and arbitrary exploration rates. *Sci Rep*, 6(37517).

Santos, F. P., Pacheco, J. M., and Santos, F. C. (2018a). Social norms of cooperation with costly reputation building. In *AAAI'2018*.

Santos, F. P., Santos, F. C., and Pacheco, J. M. (2018b). Social norm complexity and past reputations in the evolution of cooperation. *Nature*, 555:242–245.