How did human reciprocity and cooperation evolve? In A Cooperative Species, Samuel Bowles and Herbert Gintis survey the different approaches to this question, outline the evidence that bears on it, and present some of their own models of possible mechanisms. They use evolutionary game theory, replicator dynamics, and a variety of economic tools, with a mix of analysis and multi-agent simulations.

The result is a convincing argument for the importance, in the origins of altruism and reciprocity, of multi-level selection and conflict between groups, of cultural “institutions” coevolving with individual behaviours, and of social emotions and the internalisation of norms.

Bowles and Gintis begin with a survey of the experimental evidence, from behavioural economics and psychology, for altruism and other-regarding behaviours. (These are not, contrary to some presentations, irrational, but rather expressions of internalised values.) Among others, these include cooperation in prisoner’s dilemma games, “strong reciprocity” or the punishment of free riders even when that is itself costly, and retribution rather than behaviour modification (even in one-off games), sometimes extended to “antisocial punishment”.

One reason so many people have believed that humans are self-regarding, apart from economic ideology, is that it has not been clear how other-regarding behaviours could evolve. Bowles and Gintis classify the different sociobiological approaches to this, going back to work by Trivers, Wilson, and so forth. One line of explanation rests on inclusive fitness, either through kin-based selection or through multi-level (“group”) selection. Reciprocal altruism is an alternative, but while this works for dyadic interactions, with errors and imperfect knowledge it does not scale to groups of any size at all. Other possible mechanisms, resting on some kind of assortment, include indirect reciprocity with strategic reputation building and altruism as a signal of quality.

In contrast, economic thinking in this area has been dominated by “folk” theorems proving the existence of Nash equilibria in various game-theoretic models. However with imperfect public information or private information, and the possibility of error, these break down. They are “evolutionarily irrelevant equilibria”, which no group would ever find and which
would not persist even if stumbled upon. Such equilibria can work with assumptions about social norms and institutions, but a theory of origins cannot just assume the existence of these – they have to have evolved and be stably maintained, too.

A brief chapter presents some likely features of ancestral human society, drawing on ethnographic evidence from modern hunter-forager societies as well as on archaeology, ecology, and genetics. Key features here are the size of groups, the observed genetic variation between them, the frequency of inter-group conflict, and the mechanisms used to maintain social order. (One minor complaint I have here is that Pathan values are included as an example of “pre-state” norms, even though the Pathan are settled and have always been in contact with centralised states.)

With that background out of the way, Bowles and Gintis present some simulations, based on a simple selective extinction model, for the coevolution of culturally transmitted group-level institutions with individual behaviours, socially learned or genetic. For the evolution of altruism two institutions are key, “reproductive levelling” and within-group segmentation:

“if group-level institutions implementing resource sharing or positive assortment within groups are free to evolve, group-level selection processes support the coevolution of altruistic individual behaviours along with these institutions, even where these institutions impose significant costs on the groups adopting them.”

Expanding on this, parochialism and altruism probably coevolved in the context of regular war between groups, with simulation results again putting some flesh on possible mechanisms. There is also evidence that parochial altruism provides a better explanation than kin-based or reputation-based arguments for behaviour in third-party punishment and trust games, and that group conflict can stimulate altruistic punishment of free-riders. Unconditional altruism is rare: more common is strong reciprocity, with punishment of defectors. Simulations of several different models of coordinated punishment suggest that “punishers” can invade and maintain their numbers in a population: their success depends not on simple assortment but on being more likely to receive the benefits of cooperative action by others, including non-punishers. Such mechanisms work in conjunction with social emotions such as shame.

How did these evolve? Bowles and Gintis consider how a genetic predisposition to internalise norms could have evolved: there is an interplay here with the resulting norms themselves. Some simple models for gene-culture coevolution and socialization (with “oblique transmission”) suggest ways an individually fitness-reducing norm can increase the average group fitness: the cultural transmission of altruism can evolve and persist, supporting a proportion of altruists in the population.

Social emotions such as guilt and shame have the “ability to enhance the present motivational salience of future punishments” and “may function in a similar manner to pain”, averting damage, avoiding impossibly complex calculations, and helping balance impatience and other short-term goals. They are important in sustaining cooperation and a guide to how that may have evolved. Shame would have reduced the costs of strong reciprocity, with disapprobation or shaming much less costly than violence.
A brief conclusion steps back again to consider cooperation in the context of the development of linguistic capacity and cultural transmission. The social preferences that probably coevolved with our ability to cooperate remain hugely important, but Bowles and Gintis warn: “It would be wise to resist drawing strong conclusions about cooperation in the 21st century solely on the basis of our thinking about the origins of cooperation in the Late Pleistocene”.

To what extent does all this consist of “just so” stories driven by plausibility but with little empirical support – does it pass what I call “the Stoczkowski test”, after Wiktor Stoczkowski’s analysis of such stories in Explaining Human Origins? Bowles and Gintis are open about the limitations of their work: “Conclusive evidence about the origins of human cooperation will remain elusive given the paucity of the empirical record and the complexity of the dynamical processes involved. . . . the best that one can hope for is a plausible explanation consistent with the known facts.” Only in a few places do they indulge in purely speculative stories, however – for example in some off-hand, entirely incidental suggestions about the importance of projectile weapons for human evolution. Their analysis operates with a certain universality, largely abstracted from particular low-level mechanisms, and, through the use of quantitative models with explicit uncertainties in key parameters, makes the robustness of the results explicit. No radically novel ideas are introduced in A Cooperative Species. What is new in many places is the formal modelling, which provides something stable on which further work could build – even those who disagree with its conclusions may find having a concrete target useful. It is also an excellent survey of the state of knowledge in the area, as well as illustrating how ideas and tools from economics can contribute to our understanding of human evolution.