Extinction of reciprocity in a semi-automated driving coordination experiment

Keywords: experiments, social coordination, reciprocity, semi-automation, human agency

Extended Abstract

Humans face challenges in collective action when individual and group benefits are not aligned[1]. In such a situation, individuals may see no optimal way to act. However, when people have developed the norm of reciprocity, they can coordinate with others in collectively challenging situations[2]. In other words, the challenges faced in collective action can activate reciprocity among people.

Now, machine intelligence is getting involved in human social coordination and decision-making[3]. Many automation and semi-automation systems, as typified by autonomous car technology, are designed to enhance individual safety, but this could fundamentally change the structure of interactions between people. Considering the nature of collective problems and the source of reciprocity, safety support systems might actually *suppress* human sociality. However, little empirical work, if any, examines how autonomous safety control affects the norm of reciprocity in human groups. Here we test the hypothesis using a novel *hybrid-lab* experiment involving remote control robots with online participants. By combining the advantages of physical and virtual labs, in a new platform, this method allows us to systematically examine causality in collective motion with physical reality and context.

To study the emergence of reciprocity, we realized an iterated version of the game of Chicken[4], a well-established model of coordination problems, with palm-sized remote-control robots. Participants (*N*=180), recruited via Amazon Mechanical Turk, joined our game via their Internet browser from their residence. After consent, tutorials, and screening, they were randomly assigned to a pair and to one of the two robots ("yellow" and "blue" cars) that physically existed in a lab space. Participants remotely drove the assigned robot with an on-board camera view on a "single road" leading from a start grid straight to a goal area in a kind of grassland diorama (Fig. 1A). They controlled only the driving speed and whether to drive on or off the road. When they drove off the road (i.e., the outside of the printed road area), their driving speed dropped by 75%.

Participants played this remote driving game with the same counterpart over ten rounds. In each round, they were paid a decreasing bonus of up to $US1.5 depending on the speed of reaching the goal; when they did not arrive within 30 seconds, the participant earned no bonus for the round. However, the counterpart also drove their robot on the same "road" in the reverse direction. Thus, on the way to the goal, they needed to decide whether they would give way to the other by losing their time (i.e., their earnings) (Fig. 1B).

Within this basic setup, we manipulated semi-autonomous safety control systems in the robots (Fig. 1C). In the "manual" condition, participants received a warning when their car got close to an object in front. With or without warning, they needed to control their robot to avoid the obstacle (i.e., the counterpart's robot). In the "auto-braking" condition, the cars automatically stopped once in addition to the warning at the fixed distance from an object. In the "auto-steering" condition, the cars automatically swerved off-road when they went closer to the obstacle after the warning. Both players in pairs were assigned to the same condition, and they were informed of this.

Independent of the semi-autonomous systems, we also manipulated whether players were afforded the capacity to communicate. Half of the pairs could not communicate with each other. The other half had a signaling function that allowed them to send two fixed-text messages of "Go ahead." and "Thank you!" to their counterpart. The counterparts could receive the message only when they saw the sender's car in their camera view. The signaling function and limitation simulated hand signals that are often used in driving coordination. The predetermined messages did not allow players to negotiate in advance, but helped them communicate along with their behavior.

In sum, we evaluated 6 treatment combinations of semi-autonomous systems and communication capabilities. We conducted 15 sessions for each treatment combination. This experiment had a total of 90 sessions with 180 participants. Each participant played only one session consisting of 10 rounds of the remote driving game.

Our experiment shows the emergence and collapse of reciprocity with semi-autonomous systems. We categorize observed paired behavior into four types: the yellow car swerved, and the blue car went straight (unilateral turn by yellow car); the blue car swerved, and the yellow car went straight (unilateral turn by blue car); both cars swerved (bilateral turn); and they crashed. We also define the emergence of reciprocity as the temporal sequence in that players took turns giving ways across rounds.

Figure 2 shows the overall results. As the default, either of the players gave way at a rate of 29.3%, but they rarely took turns. They also had a crash at a rate of 11.3%. Auto-breaking support significantly increases unilateral turns ($p < 0.001$; penalized likelihood logistic regression), but it did not spontaneously lead to alternating reciprocity ($p = 0.209$). On the other hand, auto-steering support prevents the emergence of reciprocity completely, though it reduces accidents ($p < 0.001$). More than 90% of the time, both players leave a decision up to the machine to turn simultaneously and have no reciprocal turns in the auto-steering sessions.

Communication helps people to have alternating concessions in the anti-coordination situation, except for the sessions with auto-steering. With the signaling function, people significantly increase the reciprocal turns from 4.7% to 24.7% in the manual condition ($p < 0.001$) and from 7.3% to 29.3% in the auto-braking condition ($p < 0.001$). In the auto-steering condition, people rarely exchanged messages even when they could (only 5.3% of the time).

We found that, in contrast to the design intent, autonomous support does not simply add capabilities to humans in social coordination. Instead, it can change social factors in human behavior with the trade-off between individual safety and social reciprocity. In our experiment, the auto-braking system and the signaling functions do not avoid crashes, but encourage reciprocal concessions with communication. On the other hand, the auto-steering system takes over players’ agency in the coordination challenge. Our study suggests that whether the norm of reciprocity emerges or collapses with autonomous systems can depend on whether the technology complements or replaces human agency. Humans have developed social norms over time, but such unwritten, collective agreements could break down instantly when people leave decisions to machines.

References

1. Dawes, R. M. Social dilemmas. *Annu. Rev. Psychol.* **31**, 169–193 (1980).
2. Berg, J., Dickhaut, J. & McCabe, K. Trust, reciprocity, and social history. *Games Econ. Behav.* **10**, 122–142 (1995).
3. Rahwan, I. *et al.* Machine behaviour. *Nature* **568**, 477–486 (2019).
4. Rapoport, A. & Chammah, A. M. The game of chicken. *Am. Behav. Sci.* **10**, 10–28 (1966).

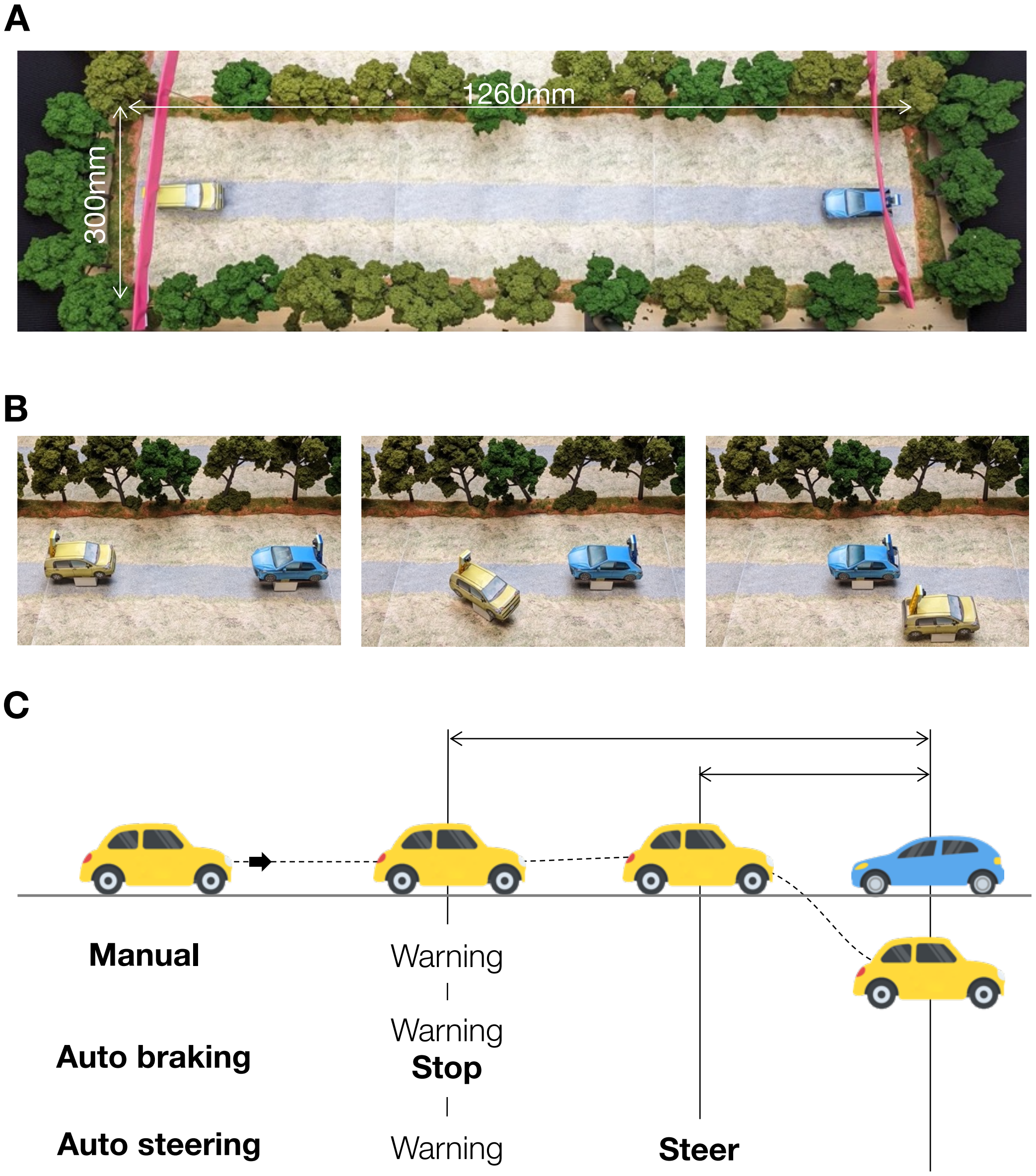

Figure 1. Experiment setup. (A) The physical coordination field. Two car robots face each other on a single road. Players remotely drive the robots over the Internet to control the speed and whether to drive on or off the road. (B) A sequence of a unilateral turn by the yellow car. To avoid a crash, at least one player needs to give way to the counterpart, but this reduces the driving speed by 75%. (C) Experimental treatments for the driving system. In addition to the default (i.e., manual driving), cars with auto-braking automatically stop once with a warning, while those with auto-steering automatically swerve at the last minute.

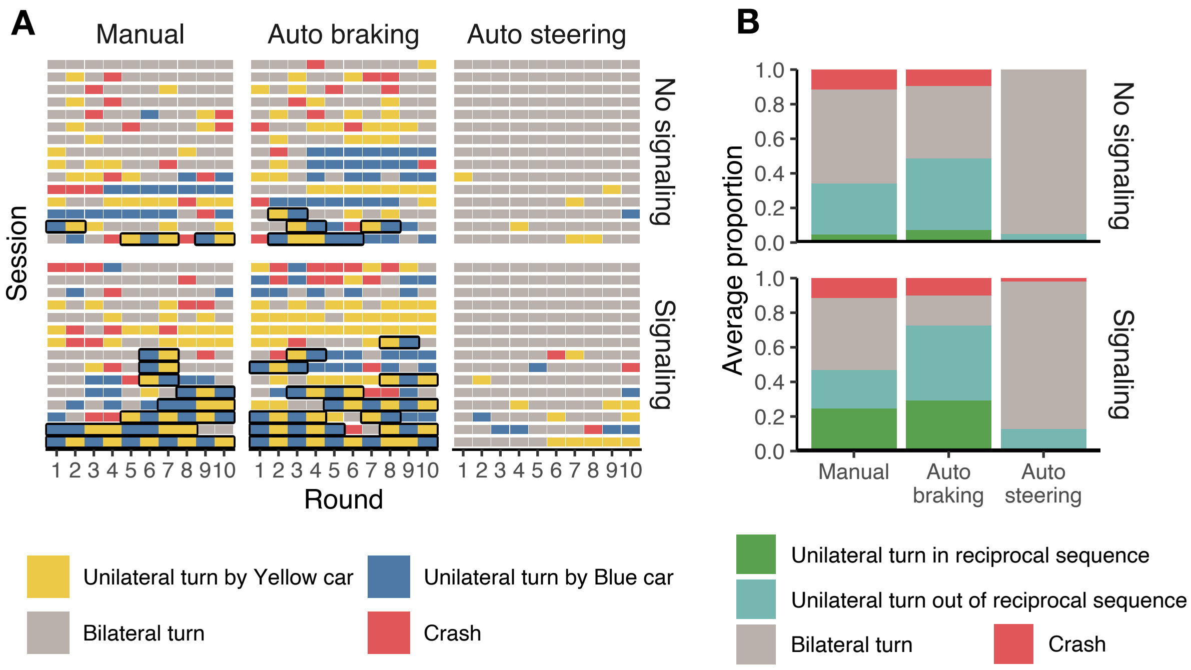

Figure 2. Experiment results (A) Paired behavior across the rounds and sessions. Each row shows a sequence of paired behavior per session. Bold outline indicates the rounds in a reciprocal sequence (B) The average proportion of paired behavior across the treatments (*N*=15 for each treatment).