PNAS RESEARCH ARTICLE | SOCIAL SCIENCES ENGINEERING OPEN ACCESS

Emergence and collapse of reciprocity in semiautomatic driving coordination experiments with humans

Hirokazu Shirado[a,1], Shunichi Kasahara[b,c], and Nicholas A. Christakis[d,e,f]

Edited by David G. Rand, Massachusetts Institute of Technology, Cambridge, MA; received May 10, 2023; accepted October 10, 2023 by Editorial Board Member Kenneth W. Wachter

Forms of both simple and complex machine intelligence are increasingly acting within human groups in order to affect collective outcomes. Considering the nature of collective action problems, however, such involvement could paradoxically and unintentionally suppress existing beneficial social norms in humans, such as those involving cooperation. Here, we test theoretical predictions about such an effect using a unique cyber-physical lab experiment where online participants (N = 300 in 150 dyads) drive robotic vehicles remotely in a coordination game. We show that autobraking assistance increases human altruism, such as giving way to others, and that communication helps people to make mutual concessions. On the other hand, autosteering assistance completely inhibits the emergence of reciprocity between people in favor of self-interest maximization. The negative social repercussions persist even after the assistance system is deactivated. Furthermore, adding communication capabilities does not relieve this inhibition of reciprocity because people rarely communicate in the presence of autosteering assistance. Our findings suggest that active safety assistance (a form of simple AI support) can alter the dynamics of social coordination between people, including by affecting the trade-off between individual safety and social reciprocity. The difference between autobraking and autosteering assistance appears to relate to whether the assistive technology supports or replaces human agency in social coordination dilemmas. Humans have developed norms of reciprocity to address collective challenges, but such tacit understandings could break down in situations where machine intelligence is involved in human decision-making without having any normative commitments.

social coordination | reciprocity | driving automation | human agency | experiments

Significance

Simple and complex forms of machine intelligence are becoming involved in many collective action challenges humans face, including ensuring safety in groups on the move. However, the social repercussions of intelligent assistance are often overlooked. We used a unique cyber-physical lab experiment involving remote-control robotic cars and widely distributed online drivers. We show that autonomous safety systems in cars can degrade the ordinary norms of reciprocity between people. Humans have developed social norms, but these can collapse when people are allowed to leave their coordination decisions to machines.

Forms of simple and complex machine intelligence are becoming increasingly involved in the collective behaviors of human groups (1–3). In particular, as a simple but important illustrative example, various active driver assistance systems are increasingly available in cars for the purpose of enhancing individual convenience and safety (4, 5). Although technical and human factors associated with such systems have been studied (6–9), the social repercussions of assistive technology have often been overlooked (10). However, because active safety assistance affects both collision risk and human agency (11), it might modify the interaction structure among people and interfere with social norms that ordinarily make self-organization possible, such as norms regarding reciprocity (12, 13). As such, the simple AI used to facilitate driving—in addition to its own importance—can provide a model for studying broader issues that arise in "hybrid systems" of humans and machines interacting in groups. Forms of AI might affect not only the behavior of the humans they interact with but might have spillover effects in how such humans interact with other humans who were not a party to the primary interaction.

Humans have developed norms of altruism and reciprocity—both evolutionarily and culturally—in order to address coordination difficulties within groups (14–16). In the case of driving cars, for instance, people often take turns giving way at intersections and junctions by tacit agreement. However, the implementation of forms of machine intelligence that affect driving might deliberately or incidentally foster or degrade natural and desirable social properties seen in self-organized collective action (17). If technology affects the normative foundations shared by a group of people, it could prompt serious or sustained social transformations and not just transient behavioral adjustments at the individual level (18, 19).

Here, we explore these matters using a game-theory model of social coordination, the chicken game (also known as the snowdrift game or the hawk-dove game) (20–22). The chicken game models two drivers, both headed down a single-lane road from opposite directions, and the first to swerve yields the road to the other (Table 1). Each driver's best outcome depends on their counterpart's choice. Thus, the decisions one expects the other to make shape one's decision (i.e., *strategic interdependence*). When both drivers wait and

Author contributions: H.S. and N.A.C. designed research; H.S. performed research; H.S. and S.K. contributed new analytic tools; H.S. analyzed data; and H.S., S.K., and N.A.C. wrote the paper and obtained funding.

The authors declare no competing interest.

This article is a PNAS Direct Submission. D.G.R. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: shirado@cmu.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2307804120/-/DCSupplemental.

Published December 11, 2023.

Table 1. Chicken game with safety assistance

		Your counterpart	
		Go straight	Swerve
You	Go straight	$0 \rightarrow b - c^*$	b
	Swerve	$b - c$	$b - c$

The quantity b indicates the maximum possible benefit when a focal actor drives the road from start to finish without obstacles. And c indicates the cost of losing time and the additional effort related to manually swerving off the road. The benefit b is larger than the cost c. When both actors seek to maximize their payoff, they will collide, which is the worst outcome (=0). When they take turns giving way, they will earn better cumulative payoffs than both swerving ($b + (b - c) > 2(b - c)$). With autosteering assistance, however, the payoff of both going straight changes from 0 to $b - c^*$, where c^* is the cost of losing time by automatically swerving. When $c^* \leq c$, each actor will keep going straight to maximize their payoff, regardless of their counterpart's choice. When $c^* > c$, actors have the same strategic interdependence as the default while they do not collide with each other.

see what the other does first, they will collide, which is the worst outcome. This coordination challenge arises in various situations while driving [e.g., crossing through an intersection, changing lanes, and merging onto an expressway (7, 8)] and in other social contexts [e.g., resource management (23) and deadlocked negotiations (24)]. Although there is no optimal solution in one-shot interactions, when the interaction happens more than once, actors can address the challenge by taking turns giving way. They are economically incentivized to practice alternating reciprocity so as to earn and share a better payoff (25, 26) (Table 1); in addition, they might see this joint problem as a social focus that gives them a reason to develop exchange relations through iterated interactions (27, 28). Thus, alternating reciprocity may emerge through locally coordinated interactions in response to the challenge of possible collisions.

However, intelligent assistance in social situations (including in automotive situations, which are used as a model system here) might affect the emergence of such reciprocity. Therefore, we explore the possible impacts of two basic active assistance systems [level 1 and 2, as defined by the Society of Automotive Engineers (29)]: autonomous emergency braking systems ("autobraking assistance") and autonomous emergency steering systems ("autosteering assistance") (Fig. 1). Autobraking assistance automatically stops a car once it detects an obstacle at a certain distance. People with autobraking assistance can thus have more time to coordinate, but they still need to control their cars to avoid collisions. On the other hand, autosteering assistance automatically steers cars to avoid obstacles; people do not need to control their cars. While autobraking assistance supports human decisions in collision avoidance as a complement to human cognition, autosteering assistance replaces human decisions and is a substitute for human cognition (30, 31).

These different associations with human agency could yield entirely different impacts on the social aspects of coordination (32). In the chicken game, autobraking assistance slows the cars facing each other on the road, which can underscore the benefit of giving way. Thus, the technology can help people to take a prosocial action. On the other hand, autosteering assistance eliminates the possibility of head-on collisions and modifies the interaction structure (Table 1). In this modified interaction structure, going straight (defection) gives individuals outcomes better than or equal to swerving (cooperation), whether the counterpart goes straight or swerves [i.e., the game is transformed into the Prisoner's Dilemma (32, 33)]. Thus, autosteering assistance can make people egocentric, i.e., prompt them to seek to maximize their payoff in a self-interested fashion by simply driving rapidly head-on. However, as a consequence, two cars with autosteering assistance heading towards each other are likely to swerve simultaneously and to rarely exchange concessions, thus reducing collective welfare (see *SI Appendix* for a more formal explanation).

Communication between people also plays a significant role in collective action (34–36). While driving, people often exchange eye contact, hand singles, and blinker signals to notify their counterparts of their intention or appreciation. If both parties share a norm of reciprocity, such signal exchanges can activate social norms and help them manifest mutual anticipation and self-organization. However, people do not obtain benefits when they do not turn communication capabilities into actions (37). Thus, communication can help people manifest reciprocity, but autosteering assistance might negate the effect because it reduces the risk of collision, which may discourage people from communicating in the first place.

We test these theoretical predictions with a unique *cyber-physical lab* experiment involving physically instantiated robotic vehicles remotely controlled by faraway online participants. By combining the advantages of physical and virtual labs (38) in a new platform, this method operationalizes the chicken game with real people in the real world to examine causality in collective behavior within a physical context (39). Participants ($N = 300$), recruited via Amazon Mechanical Turk (40, 41), joined our game via their Internet browser from their residence (*Materials and Methods*). After consent, tutorials, and screening, they were randomly assigned to a pair of people and to one of the two vehicles (a "yellow car" or a "blue car") that existed in physical space (Fig. 1*A*). Participants remotely drove the assigned vehicle with an onboard camera view on a single road leading from a start grid straight to a goal area in a kind of grassland diorama. They saw the physical space from a first-person viewpoint and controlled the driving speed and whether to drive on or off the road (*SI Appendix*). When they drove off the "road" (i.e., outside the printed road area), their driving speed dropped by 75%.

Participants played this remote driving game with the same counterpart over 10 rounds. In each round, they were paid a decreasing bonus of up to US$1.50 depending on how quickly they reached the goal; when they did not arrive within 30 s, the player earned no bonus for the round. Counterpart players drove their vehicles on the same road in the reverse direction. Thus, on the way to the goal, each one needed to decide whether to give way to the other by losing their own time and earnings (Fig. 1*B*).

Within this basic setup, we manipulated safety assistance systems in the robotic vehicles (Fig. 1*C*). In the "manual" condition, players received a warning when their car got close to an object in front. With or without warning, they needed to control their vehicle to avoid the obstacle (i.e., the counterpart's vehicle). In the "autobraking" condition, the cars automatically stopped (once) in addition to the warning at a fixed distance from an object. In the "autosteering" condition, the cars automatically swerved off-road (in opposite directions) if they went closer to the obstacle after the warning (i.e., the payoff of both going straight changed from 0 to $b - c^*$ with $c^* \leq c$ in Table 1). Finally, in the "autobraking-and-autosteering" condition, the cars were equipped with both assistance systems. The cars automatically stopped once at the same fixed distance from an obstacle; if they still moved closer to it, they reduced speed by 75% and automatically swerved off-road. A key consideration of this combined condition is the cost of going straight was higher than that of swerving (i.e., $c^* > c$ in Table 1). Thus, pairs of players could earn the most with alternating reciprocity, but they did not necessarily cooperate to avoid collisions (*SI Appendix*). Even with these assistive technologies, players needed to drive most of the time and supervise the assistive technology features (i.e., they were engaged in "semiautomated" driving). Also, autosteering assistance automatically avoided head-on collisions, but players still could collide (and some actually did; see Fig. 2) even with the assistance if they swerved from the side road to broadside their counterpart. In all

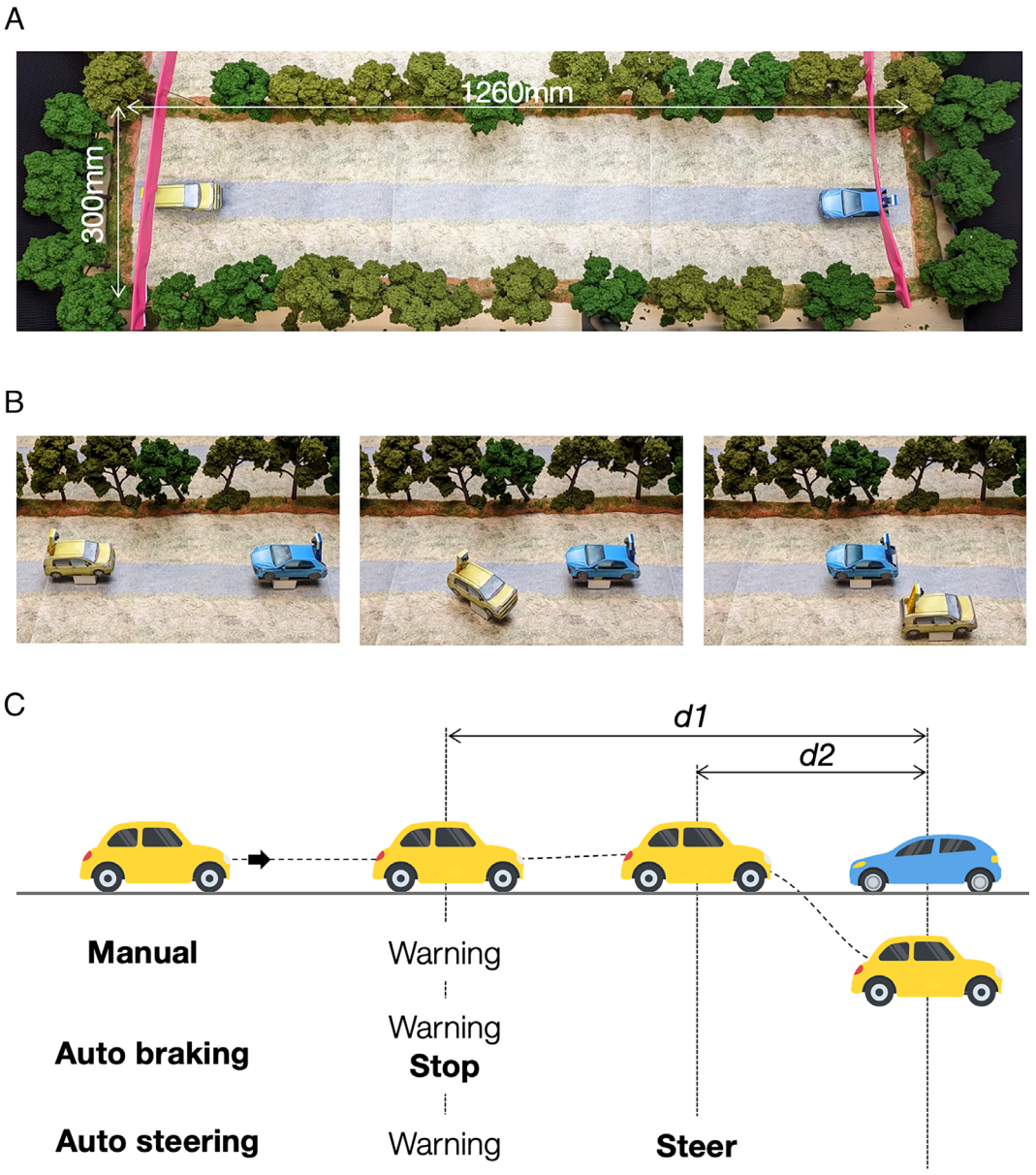

Fig. 1. Experiment setup. (*A*) The physical coordination space. Two car robots face each other on a single road. Players remotely drive the robots over the Internet to control the speed and whether to drive on or off the road. (*B*) A sequence of a unilateral turns by the yellow car. To avoid a crash, at least one of the players needs to give way to their counterpart, but this reduces their driving speed by 75% (and thus their payoff). (*C*) Experimental treatments for the driving system. In addition to the default (i.e., manual driving), cars with autobraking automatically stop once with a warning, while those with autosteering automatically swerve at the last moment.

situations, we assigned both players in pairs to the same assistive technology conditions and informed them of their own and their counterpart's assignments in advance.

Independent of the safety assistance systems, we also manipulated whether players were afforded the capacity to communicate. Half of the pairs could not communicate with each other during the game. The other half had an instant messaging function that allowed them to send two fixed-text messages of "Go ahead." and "Thank you!" to their counterpart (*SI Appendix*). Like eye contact and hand signals, the counterparts could receive the message only when they faced the sender's vehicle in their first-person camera view. The predetermined messages could help players have mutual anticipation in parallel with their actions during (but not before) the game (34, 37, 42).

Suppose intelligent assistance alters people's shared norms and not just their individual behavior. In that case, the technical impact might persist even after the assistance system stops working because people might need sufficient time to rebuild tacit collective agreements. We performed a supplementary experiment with two additional conditions to examine the persistence of any such technical impact. In the "manual to autosteering" condition, participants played the first five rounds without safety assistance (i.e., manual driving) and the last five rounds with autosteering assistance. On the other hand, in the "autosteering to manual" condition, participants played the first five rounds with autosteering assistance and the last five rounds without it. After the fifth round, we informed players of the functional change for themselves and their counterparts in these conditions. All players were able to use the messaging function in this experiment.

In sum, we evaluated 8 treatment combinations of autonomous safety technology (manual, autobraking assistance, autosteering assistance, and the combination of autobraking and autosteering assistance) and communication capabilities (presence or absence of a messaging function), and 2 supplementary conditions (manual

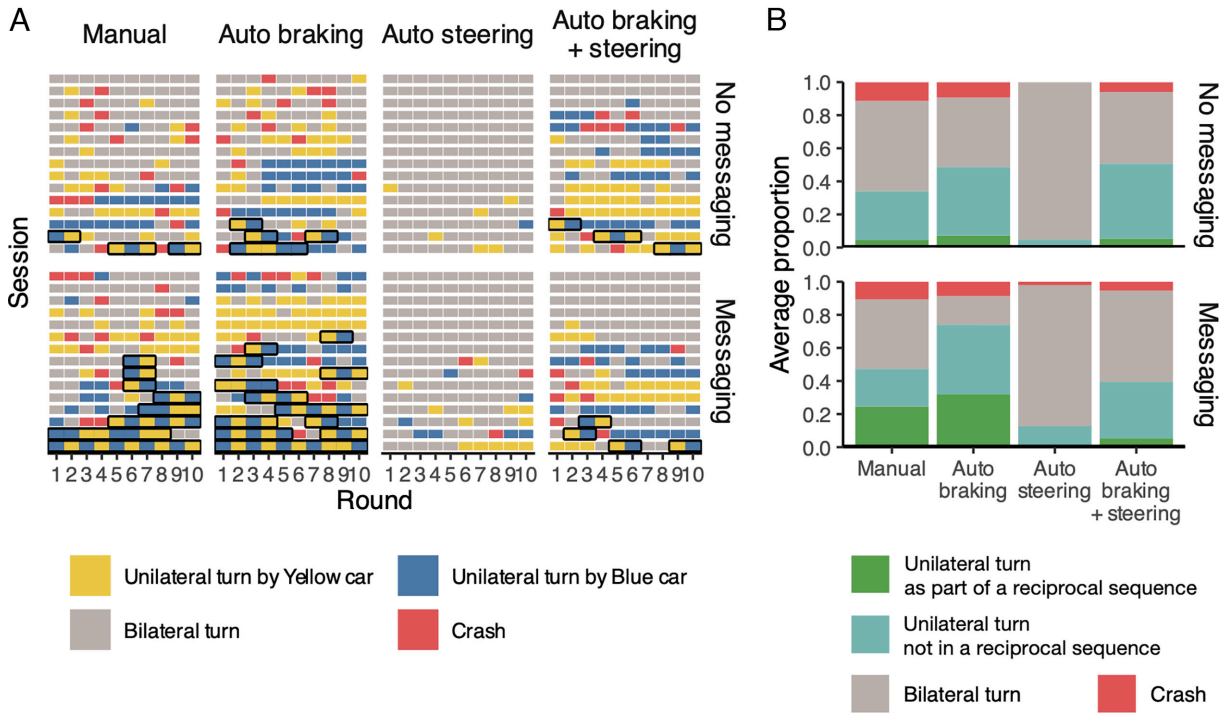

Fig. 2. Paired behavior across the conditions. (*A*) Each row shows a sequence of paired behaviors per session in eight treatment conditions. The bold outline indicates the rounds in a reciprocal sequence. (*B*) The average proportion of paired behaviors across the conditions (15 groups × 10 rounds for each condition).

to autosteering assistance and vice versa) (43). We conducted 15 sessions for each treatment combination or condition, for a total of 150 groups (sessions) with 300 participants overall (N = 240 for the main experiment and N = 60 for the supplementary one). Each participant played only one session consisting of 10 rounds of the remote driving game.

Results

Our main experiment explores the emergence and collapse of reciprocity under circumstances involving autonomous safety assistance and communication capabilities. Fig. 2 shows the overall results: We classify observed paired behavioral states into four categories: the yellow car swerved and the blue car went straight (*unilateral* turns by the yellow car); the blue car swerved and the yellow car went straight (*unilateral* turns by the blue car); both cars swerved (*bilateral* turns); and they crashed. In Fig. 2*A*, we color each session round by these four categories. We also define the emergence of reciprocity as a temporal sequence wherein players took turns giving way over multiple rounds (Movie S1). Thus, some unilateral turns were part of a reciprocal sequence, called *reciprocal* unilateral turns (otherwise, *nonreciprocal* unilateral turns), and these are indicated with bold outlines in Fig. 2*A*. Analyses evaluating whether there was evidence for exploratory behavior, such that players progressively changed over time, failed to reveal such a practice (*SI Appendix*, Fig. S1 and Table S1).

In the control group (i.e., manual driving without messaging), either of the players gave way at a rate of 34.6%, but they rarely took turns (5.3% of the time; Fig. 2*B*). They also crashed at a rate of 10.7% of the session rounds. Autobraking assistance significantly increased unilateral turns from 34.6 to 48.6% [P = 0.013; penalized multinomial logistic regression (44); *SI Appendix*, Table S2], but it neither significantly increased reciprocal turns (P = 0.169) nor significantly decreased crashes (P = 0.852). In contrast, autosteering assistance significantly reduced unilateral turns from 34.6 to 4.7% (P < 0.001), though it also reduced collisions (P = 0.005). However, no reciprocal turns emerged in the sessions with autosteering assistance (Fig. 2 *A* and *B*). Furthermore, as shown, these effects of autosteering assistance canceled out when it was combined with autobraking assistance; in the autobraking-and-autosteering condition, players behaved similarly to the autobraking condition (when they did not have the messaging function).

Communication helped people to make mutual concessions, except for the sessions with autosteering assistance. The messaging function significantly increased unilateral turns (P = 0.018), especially reciprocal ones (P < 0.001; *SI Appendix*, Table S1). Specifically, with communication capabilities, people increased reciprocal turns from 5.3 to 24.7% with manual driving and from 7.3 to 32.0% with autobraking assistance (Fig. 2*B*). However, the messaging function had little impact when people drove a vehicle with autosteering assistance. Even with communication capabilities, people still had no reciprocal turns in the autosteering condition. People also rarely reciprocated in the autobraking-and-autosteering condition although they could communicate; the fraction of reciprocal turns was 5.3%, regardless of whether players had the messaging function or not (Fig. 2 *A* and *B*).

The emergence and collapse of reciprocity are associated with corresponding changes in individual driving behavior towards a driver's counterpart (Fig. 3). Our experiment shows large variations in players' driving trajectories in the manual and autobraking conditions (Fig. 3*A* and *SI Appendix*, Fig. S2). Players gave way to their counterpart before the warning on 30.3% of occasions without messaging and 42.0% with messaging in the manual condition (Fig. 3*B*). In addition, by driving head-on, they activated the autobraking assistance at a rate of 57.0% without messaging and 54.0% with messaging (P = 0.084 and 0.804, respectively; test of equal proportions, compared with the manual condition; *SI Appendix*, Fig. S3*A*). Autosteering assistance, however, considerably reduced the individual differences in driving trajectories (Fig. 3*A*). On 97.0% of occasions without messaging and 96.0% with messaging, players did not swerve before

the warning ($P < 0.001$ for both; test of equal proportions, compared with the manual condition; Fig. 3*B*). They then left a swerving decision up to the machine at a rate of 88.3% without the messaging capability and 71.0% with it (*SI Appendix*, Fig. S3*A*). When both players relied on autosteering assistance to swerve at the last minute, they had no choice but to turn simultaneously.

Following prior work (7, 8), we apply a social value orientation (SVO) framework to players' driving behaviors to evaluate them with reference to the axes of self-interest and altruism. In this framework, SVO is represented as an angular preference ϕ that relates to how individuals weigh rewards between themselves and an alter in a coordination setting (45). We rate a player's driving behavior as rewards for themself and for their counterpart based on how far the player drives on and off the road while facing the counterpart (*Materials and Methods*). As noted above, people need to impose a cost on their counterpart in order to drive straightforward because of this game structure (Table 1). In keeping with prior empirical findings on driving coordination (7, 8), players

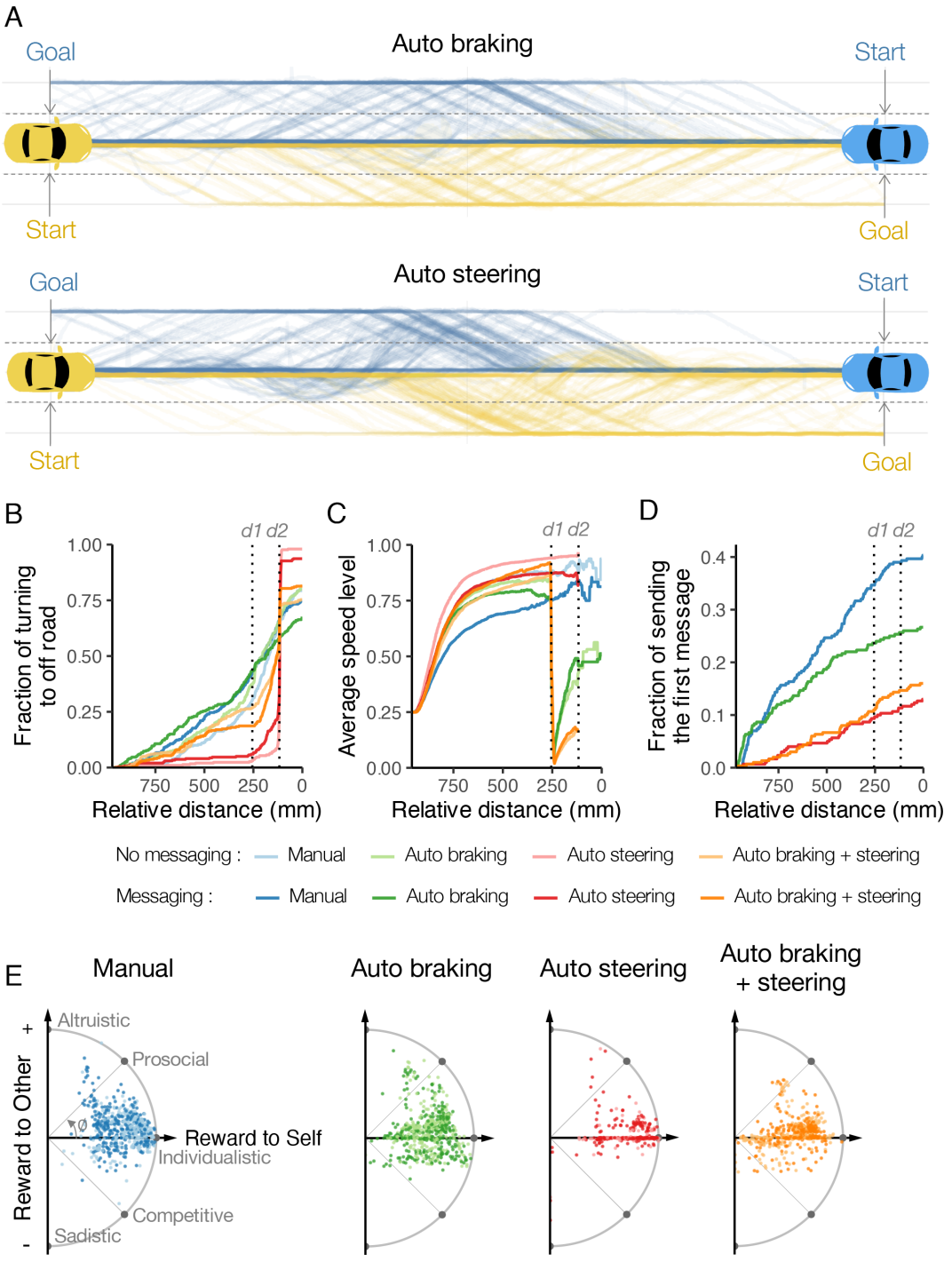

Fig. 3. Safety assistance changes human policy in driving dilemmas. (*A*) All trajectories of pairs in the autobraking and autosteering sessions with the signaling function (30 individuals × 10 rounds for each condition) (see *SI Appendix*, Fig. S2 for the other sessions). Areas between two dashed lines indicate the road in Fig. 1 *A* and *B*. The figure's horizontal to vertical ratio is the same as the experiment. (*B*) The fraction of players turning off the road, (*C*) the average speed on the road, and (*D*) the fraction of players sending the first message over relative distances across the conditions. As shown, *d1* and *d2* indicate the distances to activate warning, autobraking, and autosteering systems, as shown in Fig. 1*C*. (*E*) Player's SVO per round across the conditions ($N = 300$ for each condition; 30 individuals × 10 rounds; see *Materials and Methods*). The angular preference φ represents how individuals weigh rewards between self and others.

show various social preferences between parasitism (i.e., the "competitive" orientation in the SVO procedure) and altruism in both manual and autobraking conditions (Fig. 3*E*). Their SVO angular phases are widely distributed across individuals and rounds (*SI Appendix*, Fig. S4).

With autosteering assistance, however, these various SVOs in individuals converge into primarily self-centered ones (Fig. 3*E*). The SVO variance of the autosteering condition is significantly smaller than that of the manual condition ($P < 0.001$; *F*-test; *SI Appendix*, Fig. S4). While people behaved purely in their own interests (i.e., $-2.5° < \phi < 2.5°$) 13.6% of the time in the manual condition, they did so 79.3% of the time in the autosteering condition (from 14.4 to 88.0% without messaging; from 13.0 to 70.7% with messaging). This means that introducing autosteering assistance changed at least 65.7% of people's SVO to focus on self-interest maximization. Since we used a randomized-controlled procedure in subject recruitment, the SVO difference across the treatments cannot be explained by intrinsic individual variation.

Next, we turn to how driving behavior and communication help people achieve reciprocal coordination. Paired players needed two-step coordination to establish alternating reciprocity (*SI Appendix*, Fig. S5). First, an initiator gives way, and the other driver goes straight (*initiator's* unilateral turn). Then, on the next round, the counterpart gives way, and the initiator goes straight (*reciprocator's* unilateral turn). First, we confirmed that first-order Markov chains sufficiently represent the transitions of the paired-behavioral states across the treatments shown in our experiment (*SI Appendix*, Fig. S6 and Table S3). This indicates that each subsequent paired-behavioral state mostly depends on the immediately preceding one; thus, we did not consider two or more preceding states to estimate the state transitions.

We then examined what driving behavior and communication choice led people to take each coordination step from one state to another with logistic regression models. We find that slow, distanced swerving helps people take the first step towards prereciprocal unilateral turns, and communication helps the second step to have reciprocal unilateral turns (*SI Appendix*, Fig. S5 and Table S4).

The safety assistance systems intervene in each of these coordination steps as follows: To initiate a unilateral turn, initiators must give their counterparts enough distance and time to decide whether to go straight (*SI Appendix*, Fig. S5*A*). When a driver comes close at high speed, the counterpart anticipates that the driver will not swerve, and thus decides to swerve themself, whether the driver eventually swerves or not. In this respect, autobraking assistance can facilitate altruistic behavior (i.e., unilateral turns) by helping people to have mutual anticipation. Players significantly dropped their driving speed under autonomous braking (Fig. 3*C*), which gave them the time they needed to coordinate with each other. On the other hand, with autosteering assistance, players *increased* their driving speed and approached their oncoming counterparts at high speed (Fig. 3*C*). Thus, even when one swerved earlier than the other, it was likely too late for one's counterpart to take advantage of it.

Our analysis also shows that communication helps people reciprocate after initiating a unilateral turn (*SI Appendix*, Fig. S5*B*). When players exchanged the preformed messages—"Go ahead." and "Thank you!"—while driving, they were more likely to take turns giving way. Driving assistance technologies indirectly affect this second step by discouraging people from communicating with each other. Given the messaging function, players were more likely to send a message as they closed the distance between them—but not always. Whether to communicate depended on the assistance systems (Fig. 3*D*). In the manual condition, players sent at least one message 42.3% of the time. They reduced usage to 27.7% with the autobraking assistance ($P = 0.057$; logistic regression using the manual condition for the reference category). Autosteering assistance resulted in a further reduction to 15.3% of the time ($P = 0.004$; logistic regression). Furthermore, especially in the autosteering condition, people often missed receiving messages because they had already passed by the senders; as a result, players sent and received a message only 5.3% of the time with autosteering assistance (*SI Appendix*, Fig. S3*B*). People also communicated little with each other in the autobraking-and-autosteering condition; they used the messaging function 17.7% of the time ($P < 0.001$; logistic regression). This suppression of communication through autosteering assistance hinders the emergence of reciprocity (*SI Appendix*, Fig. S5*B*).

The behavioral changes prompted by assistance systems also affected individual performance and satisfaction. For example, people with autosteering assistance reached the goal 2.5 s faster, on average, earning US$0.125 more in every interaction than those with manual driving (*SI Appendix*, Fig. S7). Indeed, players in the autosteering condition earned more than players in any other conditions (including even the subset of players who successfully engaged in turn-taking). Although bilateral turns as a result of autosteering assistance per se are inefficient (because nobody uses a part of the main road), people improved their overall mobility efficiency with the assistive technology. On the other hand, autobraking assistance delayed people's arrival, especially when it was combined with autosteering assistance ($P < 0.001$; *SI Appendix*, Table S5). The reason is that people kept going straight more frequently when they knew they could not crash head-on into each other (Fig. 3*B*); as a result, they took more time to reach a goal due to automatic stopping and the subsequent speed reduction than manual driving.

However, in the autosteering condition, players had to give up something that they might value, namely, the opportunity to engage in reciprocity. We can estimate the value of this opportunity by comparing them to players in the manual condition. Players in the autosteering condition lost the potential of reciprocity (as shown in Fig. 2) in exchange for increasing per-round earnings from US$0.558 in the manual condition to US$0.654 in the autosteering condition (both with messaging). From an opportunity-cost perspective, in other words, players in the manual condition paid an average 14.7% "surcharge" in order to have the opportunity to engage in reciprocity with others.

Pertinently, the better economic performance with autosteering assistance did not always lead to players' subjective satisfaction in the postgame survey (*SI Appendix*, Fig. S8 and *Material and Methods*). While players' earnings significantly increased their satisfaction with themselves ($P < 0.023$), they reported no change in their satisfaction with their counterparts and cars ($P = 0.991$ and $P = 0.593$, respectively; *SI Appendix*, Table S6). Regardless of how much they earned, players were more satisfied with their counterparts when the counterparts gave way to them ($P = 0.034$) and less satisfied when they collided ($P < 0.001$). Only the messaging function increased their satisfaction with their vehicles in our model ($P < 0.001$). This gap between economic efficiency and interpersonal satisfaction aligns with theoretical and empirical findings in various social contexts (7, 45–47), showing that people have social preferences beyond economic self-interest, including preferences for interpersonal altruism, fairness, and reciprocity. Active safety assistance can undervalue such benefits in human decision-making.

Finally, we document the social inertia of autonomous assistance with a supplementary experiment where the autosteering assistance turns on or off in the middle of the game between the 5th and 6th rounds. In the manual to autosteering condition, autosteering

assistance was activated halfway. Players then quickly adjusted their coordination behavior to the functional addition (Fig. 4*A* and *SI Appendix*, Fig. S9). Using the difference-in-differences technique, we confirmed the significant effects of autosteering assistance activation on paired behavior (*SI Appendix*, Table S7). For example, players significantly reduced unilateral turns (i.e., the prerequisite for reciprocity) with the introduction of autosteering assistance, compared to the projected fraction from the results of manual condition ($P < 0.001$; Fig. 4*B*). Moreover, players significantly delayed their swerve timing, increased their driving speed, and reduced interpersonal communication in the last half of the game with autosteering assistance (*SI Appendix*, Table S8).

In contrast, players reacted sluggishly to the deactivation of assistance. In the autosteering to manual condition, assistance deactivation did not make a meaningful impact (*SI Appendix*, Table S7). For example, players did not significantly increase unilateral turns in the last half without the safety assistance, compared to the estimated fraction in case they kept using autosteering assistance ($P = 0.826$; Fig. 4*C*). Furthermore, assistance deactivation did not alter players' driving and communication behaviors (*SI Appendix*, Table S8). This contrast suggests a sustained effect of intelligent assistance. Once active assistance systems are introduced, the systems can affect the economic and normative bases for social coordination between humans, and recovery may require effort and time.

Discussion

We use the situation of intelligent assistance in driving to explore broader issues that can arise in what we have termed hybrid systems of humans and machines engaged in collective action. Our experiments here involve a basic level of driver assistance (29), i.e., technology that has become commonplace. However, recent developments in AI will enable the use of much more advanced assistance, such as fully automated driving and algorithmic traffic management (48). In a purely autonomous environment with vehicle-to-vehicle communication, each machine's decision-making can rely on a centralized control system with essentially one agent. However, the presence of human agents within the system makes interagent coordination challenging because people do not interact with machines as machines do with each other (8). Prior work has tried to address the challenge by incorporating human behavior models into the control systems of connected and automated vehicles (7, 49, 50). Our findings suggest that autonomous systems might need to further consider the specifically social and normative motives of human decision-makers to facilitate socially desirable outcomes in situations involving mixed autonomy (8, 51, 52). Automation systems in a setting of collective behavior can modify how humans treat each other, even unintentionally (53, 54).

People do not behave solely to maximize material gains (7). People often take altruistic actions and cooperate in order to align with social norms (14, 55). However, such collective understandings could break down when morality-free intelligent assistance is involved in social coordination (17). People can change their social value orientation through the presence of machine assistance, especially when it decouples the joint problem they face (27, 51).

Therefore, active safety assistance does not simply strengthen human capabilities to reduce conflicts in social coordination (31). Instead, it can also affect economic interdependence and social norms that guide individual and collective behavior with respect

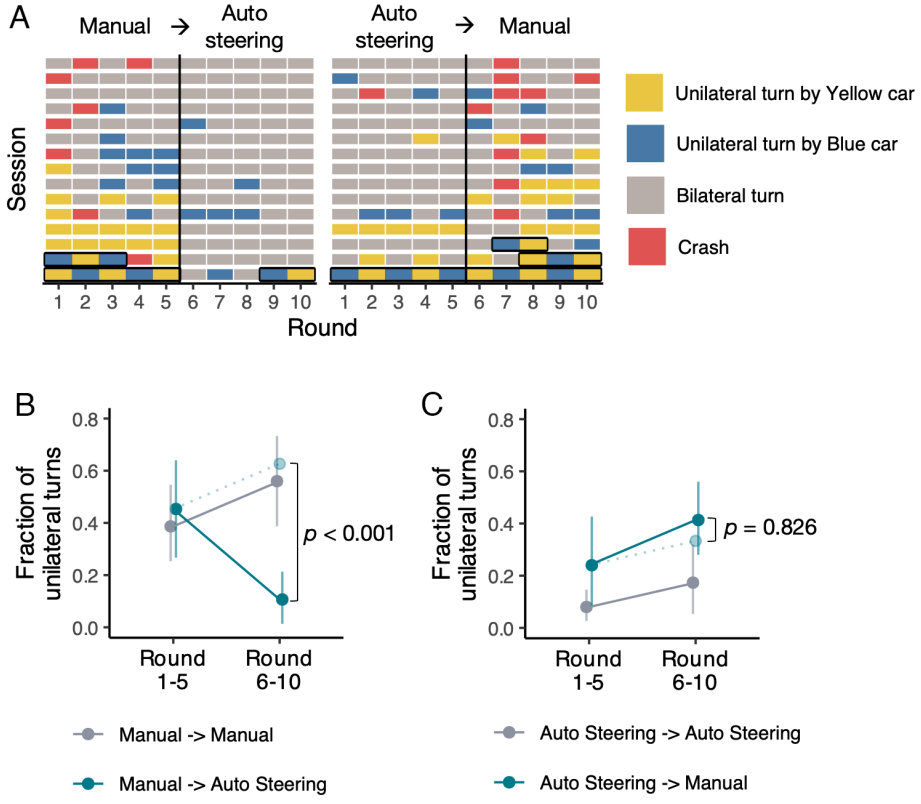

Fig. 4. Egocentric driving persists after safety assistance is lifted. (*A*) Paired behavior across the supplementary treatments ($N = 60$ individuals in 30 groups in total). Each row shows a sequence of paired behavior per session. The bold outline indicates rounds in a reciprocal sequence. The activation or deactivation of the assistive system occurs between the 5th and 6th rounds (see *SI Appendix*, Fig. S9 for the aggregate results). Changes in the fraction of unilateral turns before and after autosteering assistance activation (*B*) and deactivation (*C*), compared to the main experiment sessions without the change at the halfway mark, are shown. The translucent points indicate the counterfactual values as if the sessions had not had the change at the halfway point, given the slope in the control subjects. Error bars indicate 95% CI with bootstrapping (1,000 replications). *P* values are calculated with difference-in-differences regression models (*Materials and Methods* and *SI Appendix*, Table S7).

to trade-offs between individual interests and social reciprocity. Our experiments show that an autobraking system encourages people to make reciprocal concessions, especially in a setting affording communication capabilities. In contrast, the autosteering system significantly reduced reciprocal behavior and interpersonal communication in driving coordination, making people self-centered. People neither reciprocated nor communicated with autosteering assistance even when they could earn a better economic return from alternating reciprocity in the combined assistance condition. We also find that the adverse effects of autosteering assistance persist after the system is deactivated. These findings indicate that whether reciprocity emerges or collapses in mixed-autonomy coordination situations depends on whether machine intelligence complements or replaces human agency. Moreover, our experiment suggests two non-exclusive mechanisms for intelligent assistance to suppress human reciprocity: i) altering people's interaction structure and economic equilibrium points (56), and ii) decoupling people's sense of a joint problem, thus causing social foci to disappear (27).

Collective behavior has been studied in various contexts, such as animal migration (35, 39, 57), pedestrian flow (34, 49, 58, 59), emergency evacuation (60–62), and traffic management (63, 64). Many models of collective behavior build on dynamics involving a repulsive potential among physical particles (59, 61, 63), and some incorporate individual perception into such a physics-based approach (35, 36, 42, 49, 57). Such perception-integrated models regard individuals as actors who actively seek a beneficial destination with, or a smooth passage through, a crowd, rather than being passively repelled or propelled by others.

Our work suggests that, in addition to physical dynamics and individual perceptions, normative expectations between actors can govern collective behavior when they share and activate norms (51, 65). For example, physics-based models cannot capture the norm-referenced behaviors observed in our experiment, such as taking turns giving way over many rounds (Movie S1). On the other hand, most individuals supported by autosteering assistance behaved like physical particles, repelling each other (Movie S2), as the simple model assumes. Their bilateral turns *appear* well-coordinated from the physics-based perspective (34, 42, 59), although the actors actually had no such intention. While humans treat other people as social actors by default—which becomes especially pronounced when communication is possible—they treat others as impersonal obstacles when given active safety assistance.

Our work also introduced a unique experimental system to evaluate the actions of humans in the presence of (here, simple) AI technology in physical space, by using physical robots managed by groups of research subjects at remote distances. This model system offers advantages over having people interact solely in a virtual environment—by providing a physical instantiation of the challenges faced and by enhancing the verisimilitude of the collective action dilemmas. In fact, our simple system captured certain features of physicality, including friction and vehicular noise (39, 66). In contrast to virtual simulators (60, 67), participants behave with the understanding that their presence interacts with the real world. Reality awareness might affect individual and collective behavior in ways that simulation environments do not (68, 69). Relatedly, mathematical simulations using prearranged human models, whether based on theory or recorded data (7, 8, 48, 49, 61) also cannot capture social complexity in the same way.

Further work can examine other relevant features that we did not explore. For example, our study used an iterated two-player game to examine driving coordination. This setting is reasonable, especially in sparsely populated, closed systems (e.g., localized neighborhoods) (33). In heavily trafficked, open systems (e.g., major highways), on the other hand, people could consider more than one counterpart simultaneously to select a path (8, 34, 49, 57–59, 61). Furthermore, as people encounter more strangers there, they might make a more complicated normative commitment to coordinate with others. It is known that people can extend altruism from a specific individual to a population with indirect and generalized reciprocity (70, 71) and can develop specifically localized norms for social coordination (e.g., the "Pittsburgh left") (15, 18, 19). Moreover, people might reduce their reliance on machine intelligence according to the error rate of the system, which could lead to a different equilibrium state from our findings (72). It is therefore a worthwhile next step to confirm our findings in complex, multiagent settings.

Our study focuses on the economic payoffs associated with manual and automatic steering. We introduced a combined assistance condition to identify why drivers did not reciprocate with each other in the presence of autosteering assistance. In this condition, the economic cost of automatic steering was made higher than that of manual steering. However, it is possible that manual steering incurred additional cognitive costs to drivers, which would make the total cost of automatic steering lower than that of manual steering, still. Despite this, our results show that people in the autobraking-and-autosteering condition often made unilateral turns, similar to those in the manual condition (Fig. 2), without using autosteering (*SI Appendix*, Fig. S3*A*). This suggests that they perceived the overall cost of automatic steering to indeed be higher than that of manual driving, even when including cognitive costs. Nevertheless, they *did not* take turns giving way in the combined assistance condition, which is a clear contrast to the manual condition (Fig. 2). Further studies are needed to examine any cognitive features associated with intelligent assistance.

Although the results of laboratory experiments do not translate directly into the real world, the evidence presented here suggests that forms of AI assistance might suppress human sociality in collective behavior. People might degrade tacit agreements among themselves, such as reciprocity, in the presence of machines that can take over human agency. In the coming years, humans and machines are going to share physical space and interact "socially" in hybrid systems, such as autonomous cars sharing the road with human-driven cars. Understanding these new technologies as social catalysts among people might help us avoid any unintentional harm of intelligent automation to interpersonal interactions.

Materials and Methods

Ethics, Consent, and Preregistration. All the experiments were preregistered (*Data, Materials, and Software Availability*) and approved by the Carnegie Mellon University Committee of the Use of Human Subjects. All the participants provided informed consent. We designed the subject payments to give more than the US minimum hourly wage as of 2022. As a result, participants received US$9.25 on average (the minimum US$4.70; the maximum US$12.05) for about an 18-min task. Our data include no identifying information.

Participants. We conducted experiments from May 2022 to August 2023. A total of 300 unique participants were recruited from Amazon Mechanical Turk (MTurk) to participate in one of 150 sessions. They could not join more than one session. To minimize possible differences in driving conventions and streaming latency, we limited participants to be US-located. In addition, we required participants to be experienced workers using the MTurk qualification system. *SI Appendix*, Table S8 shows the participant demographics obtained with a free-response postgame survey. Before joining the game, all the participants passed an Internet speed check, camera visibility check, human verification check, and a comprehension test about the experiment settings, such as real-robot control and assistance capability. Actual instructions are shown in *SI Appendix*.

System. Experiments were implemented with the Breadboard platform (73). Participants interacted anonymously over the Internet using customized software playable in a browser window. The user interface has an assigned vehicle's onboard camera view of its actual environment, control buttons, and indicators for remaining time and current speed (*SI Appendix*). To avoid online participants being confused by the remote-control operation, we limited vehicles' movable area to three invisible fixed lanes: one "on-road" lane and two "off-road" lanes (Fig. 1*A*; the road image of the physical environment was for visual orientation). Thus, participants controlled their vehicles with four options: speed up, slow down, go to the right lane, and go to the left lane. They were unable to move backwards. Furthermore, we constrained leftwards and rightwards movements as follows: When participants were on the road, they could move right to avoid an obstacle. Once off the road, they could (only) move left (to return to the road). This was achieved by activating and deactivating the relevant steering buttons. When participants were assigned to a session with the messaging function, they additionally had two messaging buttons: "Go ahead." and "Thank you!" When they clicked either button, the message popped up over the current place of the sender's vehicle image in the counterpart's camera view. Thus, the sender needed to be within range of the receiver's camera for the counterpart to see the message.

Each remote-controlled vehicle consisted of an off-the-shelf robot cube [Sony's Toio (74)], a small single-board computer (Raspberry Pi Zero W), and a 120-degree-angle camera. It was covered with the paper craft of a yellow or blue car (Fig. 1 *A* and *B*). The vehicle size was about 30 × 65 × 50 mm. The robotic cubes moved at the maximum speed of about 300 mm/s and recognized their absolute location with an underbody sensor by detecting an invisible ink pattern on specific paper sheets that formed the "ground." This allowed us to control the vehicles based on their distance (e.g., for the active assistance treatments) and their location (e.g., for the 75% speed reduction during off-road driving) without relying on algorithmic location estimation. The onboard cameras were used only for participants to see the environment and control their vehicles. They faced in the direction of forward movement to show the front view with the tip of their own vehicle body (*SI Appendix*), which helped participants to sense the distance from an oncoming object. The camera view was streamed to assigned participants via the Web Real-Time Communication program (75). We confirmed that the streaming latency was small enough that people were able to control the remote vehicle with the live view in the experiment environment (mostly about 300 ms within the United States of America and less than 500 ms between the United States of America and Japan).

Procedure. After the tutorial and screening process, participants were randomly assigned to one of the eight conditions and one of the two vehicles. They practiced the remote-control operation on a separate road for two rounds. In the first practice round, they drove the assigned vehicle with no obstacles. In the second round, they had an obstacle in the middle of the road and practiced how to avoid it with autobraking or autosteering assistance when their vehicle had it. After the second round, participants were asked what obstacle they saw in the middle of the road. When both participants selected the correct answer, they were allowed to participate in the main game. This procedure guarantees all the participants of the main game had the live-streaming camera view on their browser window.

In each round, participants started on the road, facing one another in opposite directions (Fig. 1*A*). Participants were informed that both players had (or did not have) specific safety assistance technologies based on their assigned treatment. When a game round started, participants had a "Start" button to start their vehicles with the minimum speed. At the same time, a counting timer started (whether they actually started or not), showing their bonus linearly decreasing from US$1.50. When participants drove their vehicles to reach the end of the road, indicated with a pink goal tape (Fig. 1*A*), within 30 s, they would receive the indicated remaining bonus. Otherwise, they would earn nothing for the round, including when they collided. They played the game with the same counterpart over ten rounds (even after crashes). In the supplementary experiments, participants were informed of the assistance system's activation or deactivation before the 6th round started.

After the game ended, participants were asked about their driving strategy in the game, satisfaction in the game, real-world driving experience, and demographics. In the satisfaction survey, participants answered whether they were satisfied, using a 5-level rating system, with the following three statements: i) "In the game, how satisfied were you with your own drive?" ii) "In the game, how satisfied were you with your counterpart's drive?" iii) "In the game, how satisfied were you with your car and its support system?" We transformed the 5-category answers into the values from −2 to 2, with higher values indicating higher satisfaction. When they completed the postgame survey, they additionally earned the completion bonus of US$1.50 as well as base pay of US$2.00.

Analysis. We classified participants' paired behaviors into four categories: "unilateral turn by the yellow car (Y)," "unilateral turn by the blue car (B)," "bilateral turn," and "crash" (Fig. 2*A*). The classification was based on the parallel distance and intermediate point between paired cars when they went by each other (otherwise, their behavior was classified as a crash). We also identified a sequence of paired behaviors as "reciprocal" when unilateral turns alternated across the participants and rounds. Reciprocal sequences consisted of two basic patterns: Y-B and Y-Y-B-B (or B-Y and B-B-Y-Y). We did not find any case where participants reciprocated with more than two-round consecutive concessions. When we aggregated the session data to analyze paired behaviors, we used the distinction of unilateral turns based on whether the turns were a part of a reciprocal sequence instead of by which side swerved (Fig. 2*B*). We used penalized multinomial logistic regressions for analyzing the treatment effects on the multicategory paired behaviors (*SI Appendix*, Table S1). Our data shows no reciprocal unilateral turns in the sessions with autosteering assistance (Fig. 2). Logistic regression curves cannot fit such a completely separated variable across the treatments (76). Thus, we used the Firth penalization method to address the complete separation problem (44).

We used the SVO framework to evaluate each individual's driving trajectory with reference to the axes of self-interest and altruism (45) (Fig. 3*E*). Participants manifested their SVO while they interacted with each other because both payoffs to themselves and to their counterparts were contingent on their driving behavior. In contrast, their driving and proceeds in the part after they crossed were independent of their counterparts. Thus, we quantified their rewards to self and the other based on each person's travel distance towards their goal x_i until when they went by or collided at time t. Since participants earned a larger payoff when they reached the goal earlier, we defined subject i's reward to self in a round as:

$$Reward\ to\ Self_i = x_i / t.$$

When a participant swerved to give way to their counterpart, they gave a reward to their counterpart. We defined their provision as the counterpart j's travel distance during the ego's off-road movement $x_{j,freeway}$ ($x_{j,freeway} < x_j$). To calculate the reward to the other party, we excluded the case when people used autosteering assistance to swerve because they had no intention of giving rewards to their counterparts. In addition, when a participant drove on the road while their counterpart was off the road, they imposed costs on the counterpart. Thus, we defined the negative reward to the counterpart as the difference between the counterpart j's actual distance and the counterfactual one wherein the counterpart had driven on the road $x_{j,lostway}$. We calculated the counterfactual distance based on the average velocity in the last second before the counterpart swerved. Therefore, we quantified subject i's reward to the other in a round as:

$$Reward\ to\ Other_i = \left(x_{j,freeway} - x_{j,lostway}\right) / t.$$

Following the SVO framework, we evaluated each behavior's social orientation as an angular phase $\emptyset_i = \tan^{-1}(\text{Reward to Other}_i/\text{Reward to Self}_i)$ (*SI Appendix*, Fig. S3).

We used the difference-in-differences approach to evaluate the assistance activation and deactivation effects in the supplementary experiment (Fig. 4 and *SI Appendix*, Tables S6 and S7). In the supplementary sessions, autosteering assistance was activated or deactivated after the 5th round in the 10-round game. We identified the event as a treatment, compared to the control sessions without the change at the halfway point in the main experiment (i.e., the manual and autosteering conditions, respectively). We also identified the sessions from the 6th to 10th rounds as "posttreatment," compared to those from the 1st to 5th rounds. The difference-in-differences method evaluates the effect of treatment with the following regression:

$$Y = \beta_0 + \beta_1 * Treatment + \beta_2 * Post + \beta_3 * Treatment * Post + e,$$

where Y is the outcome variable (e.g., the odds of unilateral turns in Fig. 4). When the estimated interaction term's coefficient was significantly large, the activation or deactivation of assistive technology had a meaningful impact on individual or collective behavior in the middle of the game.

Data, Materials, and Software Availability. The data in this manuscript is available at Mendeley Data (43). This study was preregistered in AsPredicted.org for the main experiment (https://aspredicted.org/du7ny.pdf; https://aspredicted.org/iu6tq.pdf; https://aspredicted.org/zm3bn.pdf) and the supplementary experiment (https://aspredicted.org/g8ic6.pdf).

ACKNOWLEDGMENTS. A. Tanaka, M. McKnight, W. Israel, and K. Shimizu provided expertise useful in developing the cyber-physical experimental system. A. Zhu, E. Chien, and J. Chiu assisted in the technical implementation. L. Cheng, S. Huang, and M. Almudhry assisted in the experiment's execution. Funding for this research was provided by the Robert Wood Johnson Foundation, the NOMIS Foundation, and Japan Science and Technology (grant no. JPMJPR23I4 and JPMJPF2205).

Author affiliations: [a]Human-Computer Interaction Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15206; [b]Sony Computer Science Laboratoires, Inc., Tokyo 141-0022, Japan; [c]Okinawa Institute of Science and Technology Graduate University, Onna son, Okinawa 904-0412, Japan; [d]Yale Institute for Network Science, Yale University, New Haven, CT 06520; [e]Department of Sociology, Yale University, New Haven, CT 06520; and [f]Department of Statistics and Data Science, Yale University, New Haven, CT 06520

1. J. B. Bak-Coleman *et al.*, Stewardship of global collective behavior. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2025764118 (2021).
2. I. Rahwan *et al.*, Machine behaviour. *Nature* **568**, 477–486 (2019).
3. H. Shirado, N. A. Christakis, Locally noisy autonomous agents improve global human coordination in network experiments. *Nature* **545**, 370–374 (2017).
4. K. Bengler *et al.*, Three decades of driver assistance systems: Review and future perspectives. *IEEE Intell. Transp. Syst. Mag.* **6**, 6–22 (2014).
5. J. B. Cicchino, Effectiveness of forward collision warning and autonomous emergency braking systems in reducing front-to-rear crash rates. *Accid. Anal. Prev.* **99**, 142–152 (2017).
6. E. Yurtsever, J. Lambert, A. Carballo, K. Takeda, A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access* **8**, 58443–58469 (2020).
7. W. Schwarting, A. Pierson, J. Alonso-Mora, S. Karaman, D. Rus, Social behavior for autonomous vehicles. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 24972–24978 (2019).
8. B. Toghi, R. Valiente, D. Sadigh, R. Pedarsani, Y. P. Fallah, Social coordination and altruism in autonomous driving. *IEEE Trans. Intell. Transp. Syst.* **23**, 24791–24804 (2022).
9. C. Rödel, S. Stadler, A. Meschtscherjakov, M. Tscheligi, "Towards autonomous cars: The effect of autonomy levels on acceptance and user experience" in *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI '14* (Association for Computing Machinery, New York, 2014), pp. 1–8.
10. C. L. Bennett, E. Brady, S. M. Branham, "Interdependence as a frame for assistive technology research and design" in *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '18* (Association for Computing Machinery, New York, 2018), pp. 161–173.
11. W. Wen, Y. Kuroki, H. Asama, The sense of agency in driving automation. *Front. Psychol.* **10**, 2691 (2019).
12. M. van Veelen, J. García, D. G. Rand, M. A. Nowak, Direct reciprocity in structured populations. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 9929–9934 (2012).
13. E. Fehr, U. Fischbacher, S. Gächter, Strong reciprocity, human cooperation, and the enforcement of social norms. *Hum. Nat.* **13**, 1–25 (2002).
14. E. Ostrom, Collective action and the evolution of social norms. *J. Econ. Perspect.* **14**, 137–158 (2000).
15. S. Gavrilets, P. J. Richerson, Collective action and the evolution of social norm internalization. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 6068–6073 (2017).
16. N. A. Christakis, *Blueprint: The Evolutionary Origins of a Good Society* (Hachette UK, 2019).
17. N. Köbis, J.-F. Bonnefon, I. Rahwan, Bad machines corrupt good morals. *Nat. Hum. Behav.* **5**, 679–685 (2021).
18. R. X. D. Hawkins, N. D. Goodman, R. L. Goldstone, The emergence of social norms and conventions. *Trends Cogn. Sci.* **23**, 158–169 (2019).
19. G. Mackie, Ending footbinding and infibulation: A convention account. *Am. Sociol. Rev.* **61**, 999–1017 (1996).
20. A. Rapoport, A. M. Chammah, The game of chicken. *Am. Behav. Sci.* **10**, 10–28 (1966).
21. R. Sugden, *The Economics of Rights, Co-Operation and Welfare* (Palgrave Macmillan, London, 2005).
22. J. M. Smith, *Evolution and the Theory of Games* (Cambridge University Press, 1982).
23. K. Madani, Game theory and water resources. *J. Hydrol.* **381**, 225–238 (2010).
24. S. J. Brams, *Negotiation Games: Applying Game Theory to Bargaining and Arbitration* (Psychology Press, 2003).
25. L. Browning, A. M. Colman, Evolution of coordinated alternating reciprocity in repeated dyadic games. *J. Theor. Biol.* **229**, 549–557 (2004).
26. J. Tanimoto, H. Sagara, A study on emergence of alternating reciprocity in a 2×2 game with 2-length memory strategy. *Biosystems* **90**, 728–737 (2007).
27. S. L. Feld, The focused organization of social ties. *Am. J. Sociol.* **86**, 1015–1035 (1981).
28. R. M. Emerson, "Exchange theory, part I: A psychological basis for social exchange" in *Sociological Theories in Progress* (Houghton Mifflin, Boston, 1972).
29. On-Road Automated Driving Committee, Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles (J3016_202104, SAE International, 2021).
30. W. Nicholson, C. M. Snyder, *Microeconomic Theory: Basic Principles and Extensions* (Cengage Learning, 2012).
31. R. Parasuraman, T. B. Sheridan, C. D. Wickens, A model for types and levels of human interaction with automation. *IEEE Trans. Syst. Man Cybern. A Syst. Hum.* **30**, 286–297 (2000).
32. D. Snidal, Relative gains and the pattern of international cooperation. *Am. Polit. Sci. Rev.* **85**, 701–726 (1991).
33. R. Axelrod, W. D. Hamilton, The evolution of cooperation. *Science* **211**, 1390–1396 (1981).
34. H. Murakami, C. Feliciani, Y. Nishiyama, K. Nishinari, Mutual anticipation can contribute to self-organization in human crowds. *Sci. Adv.* **7**, eabe7758 (2021).
35. A. Strandburg-Peshkin, D. R. Farine, I. D. Couzin, M. C. Crofoot, Shared decision-making drives collective movement in wild baboons. *Science* **348**, 1358–1361 (2015).
36. M. Moussaid, S. Garnier, G. Theraulaz, D. Helbing, Collective information processing and pattern formation in swarms, flocks, and crowds. *Top. Cogn. Sci.* **1**, 469–497 (2009).
37. R. Cooper, D. V. DeJong, R. Forsythe, T. W. Ross, Communication in coordination games. *Q. J. Econ.* **107**, 739–771 (1992).
38. M. J. Salganik, *Bit By Bit* (Princeton University Press, 2018).
39. L. Li *et al.*, Vortex phase matching as a strategy for schooling in robots and in fish. *Nat. Commun.* **11**, 5408 (2020).
40. K. A. Thomas, S. Clifford, Validity and mechanical turk: An assessment of exclusion methods and interactive experiments. *Comput. Hum. Behav.* **77**, 184–197 (2017).
41. J. J. Horton, D. G. Rand, R. J. Zeckhauser, The online laboratory: Conducting experiments in a real labor market. *Exp. Econ.* **14**, 399–425 (2011).
42. M. Moussaid, D. Helbing, G. Theraulaz, How simple rules determine pedestrian behavior and crowd disasters. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 6884–6888 (2011).
43. H. Shirado, S. Kasahara, N. A. Christakis, Chicken game with safety assistance. Mendeley Data. https://data.mendeley.com/datasets/kk24ryn243/1. Deposited 7 September 2023.
44. I. Kosmidis, D. Firth, Multinomial logit bias reduction via the Poisson log-linear model. *Biometrika* **98**, 755–759 (2011).
45. R. O. Murphy, K. A. Ackermann, M. Handgraaf, Measuring social value orientation. *Judgm. Decis. Mak.* **6**, 771–781 (2011).
46. H. Shirado, G. Iosifidis, N. A. Christakis, Assortative mixing and resource inequality enhance collective welfare in sharing networks. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 22442–22444 (2019).
47. A. Nishi, H. Shirado, D. G. Rand, N. A. Christakis, Inequality and visibility of wealth in experimental social networks. *Nature* **526**, 426–429 (2015).
48. S. Feng *et al.*, Dense reinforcement learning for safety validation of autonomous vehicles. *Nature* **615**, 620–627 (2023).
49. A. Alahi *et al.*, "Social LSTM: Human trajectory prediction in crowded spaces" in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, New York, 2016), pp. 961–971.
50. D. Sadigh, S. Sastry, S. A. Seshia, A. D. Dragan, "Planning for autonomous cars that leverage effects on human actions" in *Robotics: Science and Systems* (Ann Arbor, MI, 2016), pp. 1–9.
51. M. El Zein, B. Bahrami, R. Hertwig, Shared responsibility in collective decisions. *Nat. Hum. Behav.* **3**, 1–6 (2019).
52. H. Shirado, N. A. Christakis, Network engineering using autonomous agents increases cooperation in human groups. *iScience* **23**, 101438 (2020).
53. M. L. Traeger, S. Strohkorb Sebo, M. Jung, B. Scassellati, N. A. Christakis, Vulnerable robots positively shape human conversational dynamics in a human-robot team. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 6370–6375 (2020).
54. I. Rahwan, J. W. Crandall, J.-F. Bonnefon, Intelligent machines as social catalysts. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 7555–7557 (2020).
55. J. Henrich *et al.*, In search of homo economicus: Behavioral experiments in 15 small-scale societies. *Econ. Soc. Behav.* **91**, 73–78 (2001).
56. M. A. Nowak, *Evolutionary Dynamics* (Harvard University Press, 2006).
57. S. B. Rosenthal, C. R. Twomey, A. T. Hartnett, H. S. Wu, I. D. Couzin, Revealing the hidden networks of interaction in mobile animal groups allows prediction of complex behavioral contagion. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 4690–4695 (2015).
58. J. R. G. Dyer, A. Johansson, D. Helbing, I. D. Couzin, J. Krause, Leadership, consensus decision making and collective behaviour in humans. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **364**, 781–789 (2009).
59. K. A. Bacik, B. S. Bacik, T. Rogers, Lane nucleation in complex active flows. *Science* **379**, 923–928 (2023).
60. M. Moussad *et al.*, Crowd behaviour during high-stress evacuations in an immersive virtual environment. *J. R. Soc. Interface* **13**, 20160414–20160418 (2016).
61. D. Helbing, I. Farkas, T. Vicsek, Simulating dynamical features of escape panic. *Nature* **407**, 487–490 (2000).
62. H. Shirado, Individual and collective learning in groups facing danger. *Sci. Rep.* **12**, 6210 (2022).
63. D. Helbing, B. Tilch, Generalized force model of traffic dynamics. *Phys. Rev. E* **58**, 133–138 (1998).
64. M. Bando, K. Hasebe, A. Nakayama, A. Shibata, Y. Sugiyama, Dynamical model of traffic congestion and numerical simulation. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics* **51**, 1035–1042 (1995).
65. C. Bicchieri, A. Chavez, Behaving as expected: Public information and fairness norms. *J. Behav. Decis. Mak.* **23**, 161–178 (2010).
66. P. C. Horak, J. C. Trinkle, On the similarities and differences among contact models in robot simulation. *IEEE Robot. Automation Lett.* **4**, 493–499 (2019).
67. D. Goedicke, J. Li, V. Evers, W. Ju, "VR-OOM: Virtual reality on-road driving simulation" in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18* (Association for Computing Machinery, New York, 2018), pp. 1–11.
68. P. N. Johnson-Laird, P. Legrenzi, M. S. Legrenzi, Reasoning and a sense of reality. *Br. J. Psychol.* **63**, 395–400 (1972).
69. S. H. Seo, D. Geiskkovitch, M. Nakane, C. King, J. E. Young, "Poor thing! Would You feel sorry for a simulated robot? A comparison of empathy toward a physical and a simulated robot" in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI '15* (Association for Computing Machinery, New York, 2015), pp. 125–132.
70. T. Yamagishi, T. Kiyonari, The group as the container of generalized reciprocity. *Soc. Psychol. Q.* **63**, 116–132 (2000).
71. C. Hilbe, L. Schmid, J. Tkadlec, K. Chatterjee, M. A. Nowak, Indirect reciprocity with private, noisy, and incomplete information. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 12241–12246 (2018).

72. P. Robinette, W. Li, R. Allen, A. M. Howard, A. R. Wagner, "Overtrust of robots in emergency evacuation scenarios" in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction* (Association for Computing Machinery, New York, 2016), pp. 101–108.
73. M. McKnight, N. A. Christakis, Breadboard: Software for online social experiments. https://breadboard.yale.edu/. Accessed 27 April 2022.
74. Sony Interactive Entertainment, Toy platform toio. https://toio.io/. Accessed 27 April 2022.
75. S. Loreto, S. Pietro Romano, *Real-Time Communication with WebRTC: Peer-to-Peer in the Browser* ("O'Reilly Media, Inc.", 2014).
76. M. A. Mansournia, A. Geroldinger, S. Greenland, G. Heinze, Separation in logistic regression: Causes, consequences, and control. *Am. J. Epidemiol.* **187**, 864–870 (2018).