DISCUSSION PAPER SERIES

DP13784

EFFECTIVE POLICIES AND SOCIAL NORMS IN THE PRESENCE OF DRIVERLESS CARS: THEORY AND EXPERIMENT

Antonio Cabrales, Ryan Kendall and Angel Sánchez

INDUSTRIAL ORGANIZATION, LABOUR ECONOMICS AND PUBLIC ECONOMICS

ISSN 0265-8003

EFFECTIVE POLICIES AND SOCIAL NORMS IN THE PRESENCE OF DRIVERLESS CARS: THEORY AND EXPERIMENT

Antonio Cabrales, Ryan Kendall and Angel Sánchez

Discussion Paper DP13784
Published 06 June 2019
Submitted 04 June 2019

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

EFFECTIVE POLICIES AND SOCIAL NORMS IN THE PRESENCE OF DRIVERLESS CARS: THEORY AND EXPERIMENT

Abstract

We consider a situation where driverless cars operate on the same roads as human-driven cars. What policies effectively discourage unsafe (fast) drivers in this mixed-agency environment? We develop a game theoretic model where driverless cars are the slowest and safest choice whereas faster driving speeds lead to higher potential payoffs but higher probabilities of accidents. Faster speeds also have a negative externality on the population. The model is used to create four experimental policy conditions. We fi nd that the most effective policy is a mechanism where the level of punishment (to fast drivers) is determined endogenously within the driving population.

JEL Classification: C90, D62, D63

Keywords: N/A

Antonio Cabrales - a.cabrales@ucl.ac.uk
UCL and CEPR

Ryan Kendall - ryan.kendall@ucl.ac.uk
UCL

Angel Sánchez - anxo@math.uc3m.es
Universidad Carlos III de Madrid

Acknowledgements
This research was funded by a grant from the British Academy (SRGn171072), by Ministerio de Economia y Competitividad of Spain (grant no. FIS2015-64349-P, A.S.) (MINECO/FEDER, UE) and by Ministerio de Ciencia, Innovaci?on y Universidades/FEDER (Spain/UE) (through) grant PGC2018-098186-B-I00 (BASIC).

Effective policies and social norms in the presence of driverless cars: Theory and experiment

Antonio Cabrales[*], Ryan Kendall[†], Angel Sánchez[‡§]

June 2019

Abstract

We consider a situation where driverless cars operate on the same roads as human-driven cars. What policies effectively discourage unsafe (fast) drivers in this mixed-agency environment? We develop a game theoretic model where driverless cars are the slowest and safest choice whereas faster driving speeds lead to higher potential payoffs but higher probabilities of accidents. Faster speeds also have a negative externality on the population. The model is used to create four experimental policy conditions. We find that the most effective policy is a mechanism where the level of punishment (to fast drivers) is determined endogenously within the driving population.

[*]Department of Economics, University College London, Drayton House, 30 Gordon Street, London, WC1H 0AN, United Kingdom. a.cabrales@ucl.ac.uk. Corresponding author.

[†]Department of Economics, University College London, Drayton House, 30 Gordon Street, London, WC1H 0AN, United Kingdom. ryan.kendall@ucl.ac.uk

[‡]Departamento de Matemáticas, Universidad Carlos III de Madrid; Instituto de Biocomputación y Física de Sistemas Complejos (BIFI), Universidad de Zaragoza; Unidad Mixta Interdisciplinar de Comportamiento y Complejidad Social (UMICCS), UC3M-UV-UZ; UC3M-BS Institute for Financial Big Data (IBiDat), Universidad Carlos III de Madrid

[§]This research was funded by a grant from the British Academy (SRG\171072), by Ministerio de Economía y Competitividad of Spain (grant no. FIS2015-64349-P, A.S.) (MINECO/FEDER, UE) and by Ministerio de Ciencia, Innovación y Universidades/FEDER (Spain/UE) (through) grant PGC2018-098186-B-I00 (BASIC).

1 Introduction

The World Health Organization reported 1,420,000 road traffic deaths in 2016. Furthermore, road traffic mortality is the leading cause of death for people between the ages of 15 to 29 worldwide ([26]). In the United States, the economic impact of road traffic accidents is estimated in the hundreds of billions of dollars ([5]). The OECD cites excessive speed as the number one road safety problem in most countries ([23]). A recent field experiment quantifies the relationship between speed and negative outcomes ([24]). Following a 10 mph increase in speed limits, affected freeways experienced a 3-4 mph increase in travel speed which is associated with 9-15 percent more accidents and 34-60 percent more fatal accidents ([24]). Furthermore, faster speeds have negative externalities such as elevated concentrations of carbon monoxide (14-25 percent), nitrogen oxides (9-16 percent), ozone (1-11 percent) and higher fetal death rates around the affected freeways (9 percent) ([24]).

With the hope of reducing the negative consequences of traffic, automation is poised to drastically change transportation. While currently pioneered by Tesla, even mass producing car companies such as BMW([31]), Ford ([29]), GM ([28]), and Volvo ([30]) expect fully automated models on the road by 2021. This technology will be particularly disruptive because automated models will be operating on the same roads with traditional human drivers. Furthermore, this mixed-agency driving environment will be a reality before policy makers understand, let alone implement, effective policy measures. What policies will be most effective in promoting cooperative human behavior in mixed-agency driving environments?

Policy levers that reduce driving speed and/or increase road safety in general can have immense societal benefit. This focus is related to a long tradition of studying human cooperation. Sanctions are highly effective in enforcing social norms ([11], [12], [19], [21], and [20]). A recent field experiment shows that external punishment (along with monitoring) can decrease bribing behavior in education ([6]). Developing social norms through moral suasion is also a common way to induce pro-social behavior. Another field experiment testing the policy effectiveness in the domain of energy demand shows that moral suasion and economic incentives produce substantially different policy impacts ([17]). As they show, in settings where common identity is lacking, punishment is particularly salient in encouraging cooperative behavior ([25]).

A central concern for economists and policymakers focused on driving safety is to understand what type of incentive is effective in this application. Unsurprisingly, there is a close relationship between cooperative driving behavior and exogenous punishment. For example, a 35 percent decrease in roadway troopers was accompanied with a decrease in citations and a significant increase in injuries and fatalities ([10]). In addition, congestion

tariffs in London can reduce accident rates ([14]). However, social pressures can also have an impact on driving behavior. Drivers in Tsingtao, China had less traffic violations when they received text messages with comparisons of other driving behaviors within, and outside of, the social group ([9]). Endogenous intra-group pressure can be particularly effective in enforcing a social norm. For example, study in Kenya shows that placing messages inside long-distance minibuses encouraging passengers to speak up against unsafe driving reduced insurance claims by one-half to two-thirds ([15]). Importantly, to the best of our knowledge there are no studies that address the issue of incentives and punishment in environments with sizable proportions of both human-driven and driverless cars.

In this paper, we use a theoretical and experimental approach to analyze the effect of different policies in reducing driving speeds with mixed-agency environments. This is currently the only approach available to researchers. While automation will drastically affect road safety, people driving on today's roads are only interacting with other human drivers. Therefore, we are investigating policy effectiveness of an economic environment that doesn't currently exist. While this may seem premature, policies encouraging safe driving are particularly time-sensitive. Since driving is the most deadly activity that most humans participate in, we need to understand the complex interaction between human and automated drivers before they become a reality. Society as a whole will incur a great cost if we wait for the existence of these mixed-agency environments before we start experimenting with which policies are most effective.

We develop a game theoretic model of a driving scenario where agents choose different driving speed styles, one of which is to allow their car to drive automatically. For each individual driver, faster speeds lead to higher potential payoffs and higher probabilities of being in an accident. With risk-neutral (or slightly risk-averse) drivers, the fastest driving speed is a dominant strategy. However, faster driving styles increase the probability that *all* drivers are involved in an accident. This means that faster driving styles generate a negative externality on the population, and individuals' choices can free-ride off of the safety provided by others' safer driving styles. Hence, it is possible that governmental regulation may be helpful in promoting cooperation in terms of safe driving choices.

This model is used to parametrize the *Control* condition of a laboratory experiment. The observed behavior in the *Control* condition is compared with the behavior in three treatment conditions meant to mimic possible policy interventions building on our knowledge of human cooperation. Because humans respond to framing and social comparisons, a *Framing* condition uses associative language to encourage cooperative driving behavior (this closely follows [22]). In addition, we ran two treatment conditions using punishment: *Exogenous* and *Endogenous* punishment. The *Exogenous* condition is the same as the *Control* with the ad-

dition that fast driving subjects have the possibility of incurring an exogenously determined financial penalty for choosing fast driving styles. The *Endogenous* punishment condition also includes a probabilistic fine, but the fine amount is determined by contributions by other drivers.

Our main finding is that the *Endogenous* punishment mechanism is the most effective policy to influence driving behavior, by deterring the most dangerous forms of driving. This is particularly interesting because we do not observe any change in behavior in the *Exogenous* punishment condition, despite facing the same probability of a fine from fast driving. The salience of our endogenous punishment mechanism is in alignment with findings in other problems of strategic uncertainty. For example, allowing monetary exchange systems to endogenously emerge can support a social norm of cooperation in large groups ([4] and [7]). In addition, previous research suggests that people are responsive to their "moral responsibility" in settings where each others' actions affect the population ([18]). In this previous study, as in our work, subjects are more responsive to tackling a cooperation problem themselves rather than delegating the task to an exogenous party. Also, the *Endogenous* condition combines monetary and social sanctions, which are typically salient in promoting cooperation ([21]). Finally, this result is also related to generous selling behavior in satisfaction guarantee exchange systems ([2]) as well as the importance of intra-group pressures to enforce good driving behavior ([15]). Allowing subjects to have agency in the punishment process shifts the moral responsibility to solve the problem endogenously. [1]

Interestingly, the average speed, and thus total payoffs do not change under *Endogenous* punishment. Our data suggest the following interpretation. *Endogenous* punishment reduces the number of fast drivers which, in turn, incentivizes would-be automated drivers into faster (manual) driving choices. This result is consistent with the evidence that some car safety measures do not save lives [1] because of more aggressive driving when they adjust to the security provided by seat belts or ABS brakes.

2 Theory

We model the availability of a completely autonomous car able to transport people without human intervention. Such an operation mode requires the driverless car to take actions whenever they risk colliding with another car. On a road with driverless and human cars, human drivers may free-ride off of the fact that driverless cars will prioritize safety over

[1] Although we see punishment choices decrease over time, driving behavior is consistently affected in all rounds. That is, our results are consistent with the findings that there are long-run benefits of punishment mechanisms ([13]).

speed.

What policies can mitigate this free-riding problem? To address this, we introduce a game that is a stylized representation of the problem under consideration. Key features of our model are density-dependent utility functions for both free-riding and careful drivers, including the possibility of collisions, and a formulation of the benefit in terms of travel time. In this section, we concern ourselves with the theoretical understanding of the model predictions, in order to have a proper scenario against which the experimental findings can be discussed. This framework thus opens the way to an experimental investigation of human driving behavior in the presence of driverless cars (section 3).

2.1 The game

We denote by S_i the average speed of an agent choosing driving style $i \in \{A, S, F\}$ (A stands for Automated, S stands for Slow, and F for Fast). We assume that the driving style of each action can be ordered in the following manner:

$$S_F > S_S > S_A > 0$$

If x_i denotes the proportion of type i drivers, then the Average Speed of a population is given by the following equation:

$$AS = x_F S_F + x_S S_S + x_A S_A$$

Let p_i denote the probability that a type i driver is involved in a crash.

$$p_i = a_i AS$$

We assume that the probability of a crash depends on the driving style in the following manner:

$$a_F > a_S > a_A$$

With this notation, the time needed to reach one's destination is determined by the following formulation:

$$T = \begin{cases} \frac{1}{S_i} \text{ with probability } 1 - p_i \\ \infty \text{ with probability } p_i \end{cases}$$

We can now introduce the expected utility of a driver for each driving style choice.

$$E(U(F)) = U\left(S_F\right)\left(1 - a_F AS\right), E(U(S)) = U\left(S_S\right)\left(1 - a_S AS\right), E(U(A)) = U\left(S_A\right)\left(1 - a_A AS\right)$$

The vector $x = (x_F, x_S, x_A)$ for which a driver is indifferent between choosing F, S, and A is:

$$\begin{aligned} E(U(F)) &= E(U(S)) \\ E(U(A)) &= E(U(S)) \end{aligned}$$

Assuming that all drivers share the same preferences, for a driver to be indifferent between choosing F, S, or A, the following must be true:

$$U(S_F)(1 - a_F AS) = U(S_S)(1 - a_S AS); U(S_A)(1 - a_A AS) = U(S_S)(1 - a_S AS) \quad (1)$$

Remark 1 *From equation 1 it is clear that an interior equilibrium is a solution of a linear equation system with two equations and one unknown, AS. Thus, if all players share the same preferences, an interior equilibrium occurs for a set of measure zero of the parameter values of the model.*

In order to derive experimental hypotheses, we further specify the model. The utility of drivers is a CRRA function of the inverse of the time it takes to reach one's destination.

$$u = U\left(T^{-1}\right) = T^{-\gamma}, \gamma > 0$$

This means that the expected utility of a driver for each driving style choice is

$$E(U(F)) = S_F^{\gamma}(1 - a_F AS); E(U(S)) = S_S^{\gamma}(1 - a_S AS); E(U(A)) = S_A^{\gamma}(1 - a_A AS).$$

The above model depends on the following parameters: the average speeds S_F, S_S and S_A, the crash probabilities a_F, a_S and a_A and the exponent γ in the utility function. A general analysis of the model for any value of the parameters is beyond the scope of this paper, so from now on we will focus on a set of choices for the average speeds and crash probabilities that we will later use in the experiments. This set of parameters, in which γ is still free as we cannot control risk preferences in the experiment, is as follows:

$$S_F = 2, S_S = 1, S_A = 0.5; a_F = 0.35, a_S = 0.3, a_A = 0$$

Suppose participants are heterogeneous in CRRA and γ_i follows a distribution with CDF $G(.)$. Then, we have the following

Proposition 1 *Under CRRA preferences and for our parameter values:*

1. *there are no beliefs about speed AS, and no value of* $\gamma_i \in (0,1)$ *for which it is optimal to choose* S*.*

2. *if there is a positive density of drivers for every* $\gamma_i \in (0,1)$*, there is no equilibrium where drivers only choose* A *or only* F*.*

Proof. In Appendix A. ■

Remark 1 and Proposition 1 lead to our first two hypotheses.

Hypothesis 1 *The proportion of subjects choosing* S *will be lower than those choosing* A *and* F*.*

Hypothesis 2 *Drivers in a population will never completely coordinate on choosing* A *or* F*.*

2.2 Theoretical implications of policy conditions

The game and hypotheses derived in the previous section will serve as our *Control* treatment of the experiment. The main interest of the paper is to test the effectiveness of different policy conditions (treatments) in terms of reducing the proportion of F drivers and the average speed of the population (AS). In this section, we derive theoretical results suggesting that behavior may be affected by different types of punishment (*Exogenous* and *Endogenous*) as well as the framing of the environment ($Framing$).

Exogenous (punishment). The government imposes imperfectly enforced fines for drivers choosing F. This policy imposes a (probabilistic) penalty for choosing action F, which has been shown to impact real-world driving behavior ([10] and [14]). Denote the penalty amount to be P and the probability it is imposed to be p. Then we can establish the following proposition with resulting hypothesis.

Proposition 2 *A policy using monetary punishment will decrease the proportion of drivers choosing* F *and the value of* AS*.*

Proof. In Appendix A. ■

Hypothesis 3 *The proportion of participants in the experiment choosing* F *will be lower in Exogenous than in Control.*

Framing. Some drivers who knowingly violate a social sanction (or norm) may incur a psychological cost. Such social sanctions have been shown to influence behavior in lab

settings ([22]) as well as in real-world driving environments ([9] and [15]). Suppose drivers are primed before their choice of the strategy to think that welfare of others is reduced if they choose F. Then, if they are the kind of people that suffer a cost when violating the social norm of not harming others, they would anticipate experiencing a negative utility when choosing F. Denote this disutility as P (slightly abusing notation), which makes their utility when choosing F to be

$$E(U(F)) = U\left((S_F - P)\right)\left(1 - a_F AS_i^P\right).$$

With this revised utility function, we can establish the following proposition and hypothesis

Proposition 3 *A policy that uses social sanctions will decrease the proportion of drivers choosing F and the value of AS.*

Proof. Analogous to the proof of Proposition 2 where $p = 1$ because the driver knowingly violates a social sanction. ■

Hypothesis 4 *The proportion of participants in the experiment choosing F will be lower in Framing than in Control.*

Endogenous (punishment). At a personal cost, drivers can increase the punishment cost, P, incurred by F drivers. In this way, the severity of the punishment is endogenously selected. This combination of social sanctioning along with monetary punishments has been shown to support mutual cooperation in large groups ([4], [7], [18], and [21]). In our setting, it may be in a driver's best interest to contribute to the punishment fund if they believe that it will significantly decrease the average speed of the population. Denoting the punishment as P, again slightly abusing notation, the utility of a self-interested player when choosing F would be

$$E(U(F)) = U\left((S_F - P)\right)\left(1 - a_F AS_i^P\right)$$

This revised utility function allows us to establish the following:

Proposition 4 *A policy using both monetary punishment and social sanctions will decrease the proportion of drivers choosing F and the value of AS.*

Proof. Analogous to the proof of Proposition 3. ■

Hypothesis 5 *The proportion of participants in the experiment choosing F will be lower in Endogenous than in Control.*

3 Experimental design

3.1 Participants and sessions

Experiments were conducted at a large public university. Each subject interacted in one policy condition. We conducted 8 sessions for each condition for a total of 32 experimental sessions. Each session consisted of between 8 and 12 subjects and lasted no longer than 2 hours. Appendix B further describes the data for all 326 subjects.

3.2 Task

After instructions and a test of comprehension, subjects interacted in a multi-round decision-task. The number of rounds was randomly determined to be between 17 and 25 and the subjects did not know which period would be the final one in their experiment.[2] In each round, subjects made two incentivized choices - (1) a driving style choice and (2) a guess about the driving style choices of other participants in the room. The remainder of this subsection describes the choice environment that is the same across policy conditions. Screen shots for all conditions are in Appendix D.

In each round, every subject chose whether to drive Fast (F), Slow (S), or Auto (A). The payoffs for each choice were consistent with the parametrization described in the previous section. Because subjects were paid for one randomly selected round, the payoffs were scaled (by 14). In this way, payoffs were represented as GBP during the task. Thus, conditional on not being in an accident in a given round, the subjects who chose F, S, and A earned £28, £14, and £7, respectively. In addition, the probabilities of being in an accident were $a_F = 0.35, a_S = 0.3, a_A = 0$ times the average speed, AS.

In each round, every subject submits a guess about the proportion of subjects in the room who would choose F, S, and A. They did so by using the computerized "triangle tool" which allowed subjects to make their guess by dragging a point within a triangle where each vertex of the triangle represented a guess where 100% of the subjects in the room were choosing one driving style. The amount a subject earned from their guess was £5 minus the difference between their guessed distribution of driving styles and the actual distribution of driving styles in that round. A perfect guess would earn £5 and a very inaccurate guess would earn £0.

The triangle tool was also used by subjects to calculate the probability of an accident for

[2]Starting in round 18, there was a $\frac{2}{3}$ chance that another round would be played. This process continued until round 25 was reached, which was determined to be the last round. Subjects were told that "The experiment will last between 18 and 25 rounds. The exact number of rounds is randomly determined by the computer." A computer error stopped one session in round 17 instead of round 18.

each driving style conditional on a possible distribution of driving styles. The probability of being in an accident (and earning £0) for each driving style was updated when the subject changed their guess about the population. This way, subjects could compare the probabilities of accidents for different driving style choices when facing different beliefs about the distribution of drivers in the population.

Starting in round 2, subjects had complete information about their choices, the choices of other participants in the room, and their payoffs in all previous rounds. In addition, a picture was shown in the top-left of the screen which showed the distribution of driving style choices in the previous round as well as that subject's guess about the distribution in the previous round.

After every subject submitted their driving choice and their guess about the distribution of the other subjects in the room, they were shown a results screen summarizing the past round. This screen showed the subject's earnings based on the accuracy of their guess about the population. In addition, each subject was informed about their probability of being in an accident, the realization of this event, as well as their total payoff from their driving choice.

3.3 Policy conditions

Control. Subjects participated in the experiment described above. Subjects chose between driving "Fast", "Slow", or "Auto" and were incentivized to guess the distribution of these driving types within the "population" of other subjects.

Framing. Subjects chose between driving "Reckless", "Slow", or "Safe" and were incentivized to guess the distribution of these driving types within the "community" of other subjects. This type of associative framing can increase contribution rates in public goods games ([22]).

Exogenous (punishment). Subjects who chose F had a 25% chance to pay a fine of £4. This fine only applied to subjects who were not in an accident in that round.

Endogenous (punishment). Subjects who chose F had a 25% chance to pay a fine of £X. X was determined every round in the following way. When subjects were making a driving style choice and their guess about the population, they also had to choose whether to contribute £1 into a fund used to punish F drivers. The fine amount (X) equals the number of subjects who contributed to the punishment fund times 2.5. This fine only applied to subjects who were not in an accident in that round.

4 Results

4.1 Driving style choices

How do the policies affect driving style choices? Table 1 shows the percentage of driving style choices observed across all rounds separated by condition. Overall, the profile of driving choices in *Framing* and *Exogenous* are not significantly different from *Control* (Pearson's Chi-Squared p-value = 0.665 and 0.144, respectively). *Endogenous* shows the largest effect on behavior with a profile of choices that is significantly different from *Control* at the 0.016 level. *Endogenous* has a decrease in the slowest (Auto) *and* fastest drivers along with an increase in the moderate driving style choice (Slow).

Table 1: Choices by condition

Policy	% Fast	% Slow	% Auto
Control	46.69%	19.71%	33.60%
Framing	45.98%	21.00%	33.01%
Exogenous	48.74%	20.79%	30.48%
Endogenous	44.50%	23.69%	31.82%

To further explore the two-level effect of the *Endogenous* condition, we use a multinomial logistic regression. This will examine the relationship between the (nominal) driving style choice and the effect of the three policy conditions relative to *Control*. This process conducts 2 independent binary logistic regressions in which one driving style is used as a reference which the other 2 driving styles are regressed against. For expositional clarity, we designate the Slow driving style as the reference. In doing so, the model is expressed by the following two equations:

$$\ln\left(\frac{prob(Auto)}{prob(Slow)}\right) = Framing \cdot \beta_{1,A} + Exogenous \cdot \beta_{2,A} + Endogenous \cdot \beta_{3,A} + \mathbf{X}\beta_{\mathbf{X},\mathbf{A}} + \beta_{0,A}$$

$$\ln\left(\frac{prob(Fast)}{prob(Slow)}\right) = Framing \cdot \beta_{1,F} + Exogenous \cdot \beta_{2,F} + Endogenous \cdot \beta_{3,F} + \mathbf{X}\beta_{\mathbf{X},\mathbf{F}} + \beta_{0,F}$$

$\mathbf{X}$ is a vector of subject-specific and period-specific control variables. We control for the effect of earnings and accidents in the proceeding round ("PrevEarn" is an integer between

-1 and 28, "PrevAcc"= 1 if the subject experienced an accident in the previous round) as well as a dummy variable tracking early and late rounds ("LateRounds"= 1 in rounds 11 and greater). These variables may be important if subjects are learning to play the game differently over time. After the choice periods, we collected data on subject-specific variables. Subjects made incentivized decisions in a multiple-price list to elicit risk preferences ([16];"Risk" $\in [0, 10]$ where 10 is very risk-loving). Subjects were asked their sex ("Sex" = 1 if male), whether they had ever been issued a driver's license ("Driving" = 1 if yes), and whether they felt they became better at earning money during the experiment ("Learning" = 1 if yes).

Maximum likelihood estimates are shown in Table 2.[3] Our main coefficients of interest are connected to the dummy variables for each policy. These estimated coefficients show the log odds that a certain driving style (either Auto or Fast) is chosen relative to the reference driving style (Slow). For example, consider the *Endogenous* coefficient in column 1a of Table 2 (-0.238). This means that the log odds of choosing Auto as opposed to Slow is decreased by 0.238 when comparing driving behavior in *Control* to driving behavior in *Endogenous*. We fit one model using only the policy variables (columns 1a and 1b), one model including our control variables (2a and 2b), and one model including our control variables as well as interactions between Risk and Sex (3a and 3b).

[3]Subject-level fixed effects are not included because each subject experiences only one condition.

Table 2: The effect of policy conditions on driving choices (relative to Slow)

	Auto	Fast	Auto	Fast	Auto	Fast
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)
Framing	−0.081	−0.079	−0.065	−0.090	−0.213	−0.901***
	p = 0.418	p = 0.406	p = 0.528	p = 0.358	p = 0.461	p = 0.002
Exogenous	−0.151	−0.010	−0.048	0.019	0.186	−0.919***
	p = 0.120	p = 0.912	p = 0.634	p = 0.843	p = 0.517	p = 0.002
Endogenous	−0.238**	−0.231***	−0.275***	−0.196**	−0.879***	−1.738***
	p = 0.011	p = 0.009	p = 0.005	p = 0.031	p = 0.002	p < 0.001
Risk			−0.247***	−0.017	−0.298***	−0.158***
			p < 0.001	p = 0.420	p < 0.001	p < 0.001
Sex			0.195***	0.570***	0.412***	0.162
			p = 0.007	p< 0.001	p = 0.004	p = 0.227
PrevEarn			−0.045***	0.065***	−0.044***	0.064***
			p < 0.001	p < 0.001	p< 0.001	p< 0.001
PrevAcc			−0.959***	1.374***	−0.936***	1.343***
			p < 0.001	p < 0.001	p < 0.001	p < 0.001
Learning			−0.223***	0.057	−0.276***	0.025
			p = 0.003	p = 0.401	p < 0.001	p = 0.717
Driving			0.291***	0.071	0.251***	0.041
			p < 0.001	p = 0.309	p = 0.002	p = 0.563
LateRounds			0.397***	0.145**	0.396***	0.147**
			p < 0.001	p = 0.029	p < 0.001	p = 0.027
*Framing**Risk					0.116*	0.141**
					p = 0.063	p = 0.019
*Framing**Sex					−0.690***	0.418**
					p = 0.001	p = 0.034
*Exogenous**Risk					−0.021	0.190***
					p = 0.737	p = 0.002
*Exogenous**Sex					−0.429**	0.211
					p = 0.042	p = 0.272
*Endogenous**Risk					0.144**	0.279***
					p = 0.027	p < 0.001
*Endogenous**Sex					0.143	0.917***
					p = 0.468	p < 0.001
Constant	0.533***	0.862***	1.687***	−0.642***	1.843***	0.202
	p < 0.001	p < 0.001	p < 0.001	p < 0.001	p < 0.001	p = 0.352
Akaike Inf. Crit.	14,185	14,185	12,705	12,705	12,633	12,633

Note: *p<0.1; **p<0.05; ***p<0.01

Endogenous has a consistent and significant effect on driving behavior in all 3 models. Both Fast and Auto drivers transition into Slow as we compare *Control* with *Endogenous*.

Result 1 *Compared to Control, subjects in Endogenous are more likely to choose Slow than Auto. In addition, subjects are more likely to choose Slow than Fast.*

This result suggests that the *Endogenous* policy is effective at deterring the worst kind of behavior (fast drivers). However, in doing so, *Endogenous* also promotes slightly more dangerous behavior from people who were originally avoiding it because of the previous higher presence of dangerous drivers. As mentioned in the introduction, this is reminiscent of the evidence that seat belts do not save lives [1] because drivers adjust their behavior to the higher safety awarded by the seat belts.

A superficial reading of column 3b could suggest that *Framing* and *Exogenous* may encourage Fast drivers into Slow drivers. However, the other two models appear inconsistent with this result. This is due to the fact that model 3b includes interactions of different variables with the condition, and hence those coefficients should be read as the impact of the condition on the excluded category. So the correct interpretation is that the average treatment effect of *Framing* and *Exogenous* is null in terms of turning Fast into Slow drivers, but that masks some positive and some negative interactions with demographic characteristics, which we discuss below. Furthermore, *Framing* and *Exogenous* do not influence drivers who, in the *Control* environment would choose Auto.

It is surprising to notice that *Exogenous* and *Endogenous* have drastically different impacts on driving behavior. Both policies are focused on promoting cooperation through the punishment of Fast drivers. The only difference was that the amount of the punishment was decided by costly investment within the population. This result mimics the intuition behind previous findings which demonstrate how social norms are more deeply incorporated when they are formed endogenously ([2], [4], [7], [13], [15], [18], [21]).

Some estimates of control variables are of interest when they are consistent across models 2 and 3. Subjects more likely to choose Auto (compared to Slow) are more risk-averse, male, owned a driver's license, and did not get better at earning points by the end of the experiment. In addition, Auto was preferred to Slow when a subject earned less and did not have an accident in the preceding period, and during the later rounds of the experiment. Fast was preferred to Slow when a subject earned more in the preceding period and during the later rounds of the experiment.

The interaction variables from model 3 are also interesting. The effects of *Framing* and *Exogenous* on driving behavior were dependent on a subject's sex. For instance, under these policies, males were more likely to choose Fast and less likely to choose Auto when

compared to their female counterparts. In *Endogenous*, males were more likely to choose Fast. Furthermore, all three policies seemed to have a larger effect on subjects with higher risk aversion.

4.2 Average Speed

The individual driving style choices determine the population's Average Speed. This is an important outcome because it determines the probability of an accident for the Fast and Slow drivers. More generally, Average Speed measures the general safety of a driving environment. *How do the policies affect a population's average speed?*

A linear regression studies the relationship between Average Speed and the effect of the three policy conditions. In line with Table 2, we also control for a range of other variables. Unlike in Table 2, Average Speed is a group-level outcome, which means the control variables are aggregated to reflect the group-level values, rather than individual-level values. "Risk (Avg)" is the average of a group's "Risk" measure used in Table 2. Similarly, "Sex (Prop)" is the proportion males in a group. Average earnings and proportion of accidents in proceeding rounds are also controlled for ("PrevEarn (Avg)" and "PrevAcc (Prop)"). The proportion of subjects in a group who have been issued a driver's license and who felt they learned during the experiment were controlled for ("Driving (Prop)" and "Learning (Prop)"). Finally, a dummy variable tracks early and late rounds ("LateRounds"= 1 in rounds 11 and greater).

Estimates are shown in Table 3. Our main coefficients of interest are connected to the dummy variables for each policy. We fit one model using only the policy variables (column 1), models including all control variables and all interaction variables between Risk (Avg) and Sex (Prop) (2a and 3a), and models including only the control variables as well as interactions between Risk (Avg) and Sex (Prop) (2b and 3b) that provide the best fit of the data (according to the AIC).

Framing and *Exogenous* have no significant effect on Average Speed in any of the models. The *Endogenous* policy only shows a reduction in Average Speed with marginal significance in model 3b (p-value 0.093). As for controls, a higher average risk preference, lower proportion of males, and a higher proportion of accidents in the previous round increases Average Speed. In addition, groups with high proportions of males significantly increased Average Speed in the *Exogenous* condition.

Result 2 *None of the treatments have a robust significant effect on Average Speeds compared to the Control groups.*

Table 3: The effect of policy conditions on Average Speed

	(1)	(2a)	(2b)	(3a)	(3b)
Framing	0.010	0.005	0.007	−0.017	0.022
	p = 0.617	p = 0.831	p = 0.725	p = 0.938	p = 0.915
Exogenous	0.051**	0.034	0.030	−0.121	−0.108
	p = 0.013	p = 0.101	p = 0.142	p = 0.469	p = 0.506
Endogenous	0.004	0.013	0.003	−0.280	−0.303*
	p = 0.844	p = 0.520	p = 0.863	p = 0.125	p = 0.093
Risk (Avg)		0.050***	0.056***	0.060**	0.066**
		p< 0.001	p< 0.001	p = 0.030	p = 0.016
Sex (Prop)		−0.026	−0.039	−0.235***	−0.250***
		p = 0.590	p = 0.403	p = 0.003	p = 0.002
PrevAcc (Prop)		0.126***	0.113***	0.102**	0.092**
		p = 0.005	p = 0.006	p = 0.020	p = 0.024
PrevEarn (Avg)		0.002		0.001	
		p = 0.373		p = 0.554	
Learning (Prop)		−0.110**		−0.068	
		p = 0.045		p = 0.227	
Driving (Prop)		−0.030		0.008	
		p = 0.586		p = 0.898	
LateRounds		−0.004		−0.003	
		p = 0.777		p = 0.810	
*Framing**Risk (Avg)				−0.018	−0.027
				p = 0.662	p = 0.499
*Framing**Sex (Prop)				0.184	0.183
				p = 0.186	p = 0.185
*Exogenous**Risk (Avg)				−0.021	−0.026
				p = 0.549	p = 0.454
*Exogenous**Sex (Prop)				0.517***	0.524***
				p< 0.001	p< 0.001
*Endogenous**Risk (Avg)				0.057	0.062
				p = 0.184	p = 0.141
*Endogenous**Sex (Prop)				0.113	0.107
				p = 0.401	p = 0.424
Constant	1.287***	1.133***	1.051***	1.165***	1.128***
	p< 0.001	p< 0.001	p< 0.001	p< 0.001	p< 0.001
Obs.	660	660	660	660	660
R^2	0.012	0.069	0.061	0.104	0.102
Adj. R^2	0.007	0.055	0.053	0.082	0.085
Res. SE	0.186	0.182	0.182	0.179	0.179
F Stat	2.592*	4.809***	7.124***	4.688***	6.117***
AIC	-341	-366	-369	-380	-386

Note: *p<0.1; **p<0.05; ***p<0.01

4.3 Beliefs about the driving choices within the population

How do the policies affect a subject's beliefs about the driving style choices of others? Three separate linear regressions study the relationship between the three policies and a subject's belief about the choices of other subjects in the population. As shown in Table 4, this produces one model estimating the effect on a subject's believed proportion of Fast drivers (1a and 1b), Slow drivers (2a and 2b), and Auto drivers (3a and 3b).[4]

These estimated models suggest two findings. First, subjects in the *Endogenous* condition accurately predict the change in driving behavior relative to the Control condition. The beliefs in *Endogenous* are in the exact same direction as the choices - less Fast and Auto drivers and more Slow drivers. Second, subjects in the *Framing* and *Exogenous* conditions believe that these policies would increase the number of Fast drivers and decrease the number of Auto drivers (with no effect on the number of Slow drivers). However, these beliefs are inconsistent with Table 2, which suggests that these conditions have no impact on driving behavior on average.

Result 3 *Unlike the subjects in Framing and Exogenous, subjects in Endogenous accurately predict the effect of the policy on the driving behavior of others.*

4.4 The relationship between beliefs and driving choices

How does a subject's beliefs about the population determine that subject's driving choice? Four separate linear regressions (one for each experimental condition) study the relationship between a subject's belief about the choices of other subjects in the population and that subject's belief. The results of these 4 regressions are shown in Table 5 which omits the estimated coefficients for the control variables.

Result 4 *In all 4 experimental conditions, if a subject believes there will be more Slow drivers, s/he is more likely to drive Slow as well.*

Result 4 suggests that the Slow action seems to be an attractive socially reinforcing action or as an action with strategic complementarity.

In *Control*, an increase in a subject's belief about the proportion of Fast drivers corresponds with that subject choosing Auto or Fast more likely than Slow. This can be explained as subjects reacting to two tendencies; one to be strategic and one to follow the norm of the population. Subjects who choose Auto when their Fast belief is high are playing the game as game theorists would predict whereas subjects who choose Fast when their Fast

[4] A Tobit model produces similar results.

Table 4: The effect of policy conditions on beliefs about other drivers

	Fast Belief	Slow Belief	Auto Belief	Fast Belief	Slow Belief	Auto Belief
	(1a)	(2a)	(3a)	(1b)	(2b)	(2c)
Framing	1.368***	0.380	−1.748***	1.480***	0.025	−1.505***
	p = 0.004	p = 0.412	p = 0.001	p = 0.002	p = 0.956	p = 0.003
Exogenous	3.400***	0.474	−3.874***	3.293***	0.025	−3.318***
	p < 0.001	p = 0.286	p < 0.001	p < 0.001	p = 0.954	p < 0.001
Endogenous	−1.213***	3.443***	−2.230***	−1.087**	3.381***	−2.293***
	p = 0.006	p < 0.001	p < 0.001	p = 0.012	p < 0.001	p < 0.001
Risk				0.464***	0.382***	−0.846***
				p < 0.001	p < 0.001	p < 0.001
Sex				0.741**	−1.375***	0.634*
				p = 0.020	p < 0.001	p = 0.065
PrevEarn				0.210***	−0.069***	−0.141***
				p < 0.001	p < 0.001	p < 0.001
PrevAcc				3.662***	0.237	−3.899***
				p < 0.001	p = 0.594	p < 0.001
Learning				0.799**	−0.321	−0.478
				p = 0.015	p = 0.316	p = 0.175
Driving				0.133	−0.130	−0.003
				p = 0.694	p = 0.696	p = 0.993
LateRounds				2.862***	−5.628***	2.766***
				p < 0.001	p < 0.001	p < 0.001
Constant	44.090***	22.156***	33.754***	36.729***	25.220***	38.051***
	p < 0.001	p < 0.001	p < 0.001	p < 0.001	p < 0.001	p < 0.001
Obs.	6,749	6,749	6,749	6,749	6,749	6,749
R^2	0.017	0.012	0.010	0.058	0.068	0.041
Adj. R^2	0.017	0.011	0.009	0.056	0.067	0.040
Res. SE	13.166	13.014	14.142	12.899	12.645	13.921
F Stat.	39.625***	26.926***	21.843***	41.240***	49.189***	29.022***

Note: *p<0.1; **p<0.05; ***p<0.01

belief is high are playing the game as if they are following a social norm. However, only the strategic tendency remains strong in the 3 policy conditions, as an increase in a subject's belief about Fast drivers is associated only with more Auto choices. This effect is strongest in the *Endogenous* condition.

A similar analysis is true for a subject's increasing belief in the proportion of Auto drivers. However, both tendencies are observed in the *Control* and *Framing* condition, whereas only the social norm tendency is present in the two punishment conditions.

Table 5: The effect of beliefs on driving choices (relative to Slow)

	Control		*Exogenous*		*Framing*		*Endogenous*	
	Auto	Fast	Auto	Fast	Auto	Fast	Auto	Fast
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Fast Belief	0.037^{***} $p < 0.001$	0.019^{***} $p < 0.001$	0.029^{***} $p < 0.001$	0.008^{**} $p = 0.046$	0.025^{***} $p < 0.001$	0.007 $p = 0.138$	0.025^{***} $p < 0.001$	-0.004 $p = 0.301$
Slow Belief	-0.019^{***} $p < 0.001$	-0.032^{***} $p < 0.001$	-0.014^{***} $p = 0.006$	-0.038^{***} $p < 0.001$	-0.028^{***} $p < 0.001$	-0.042^{***} $p < 0.001$	-0.033^{***} $p < 0.001$	-0.047^{***} $p < 0.001$
Auto Belief	0.035^{***} $p < 0.001$	0.008^{*} $p = 0.059$	0.054^{***} $p < 0.001$	0.004 $p = 0.271$	0.053^{***} $p < 0.001$	0.016^{***} $p = 0.002$	0.034^{***} $p < 0.001$	-0.004 $p = 0.319$
Const.	0.00^{***} $p < 0.001$	0.00 $p = 0.492$	0.00^{***} $p < 0.001$	-0.00^{***} $p = 0.001$	0.00^{***} $p < 0.001$	0.00 $p = 0.015$	0.00^{***} $p < 0.001$	-0.00^{***} $p < 0.001$
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
AIC	3,142	3,142	2,938	2,938	2,618	2,618	3,391	3,391

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

5 Discussion and conclusion

The inevitable introduction of driverless cars raises important economic and social concerns over road safety. Adoption of this technology will be gradual which means that we will need regulatory policies for mixed-agency driving environments with both human-driven and driverless cars. In this paper, we have made a first attempt to study which policies can reduce driving speeds - thus increasing road safety. We have proposed three different interventions, namely framing the situation in a safety-conscious manner, putting in place fines by an external agent (traffic police), and setting up a scheme in which fines are imposed according to the participation of the drivers' community. Of these three scenarios, we have found that endogenizing the punishment mechanism is the only way to influence cooperative behavior in our novel mixed-agency driving scenario.

As far as the effectiveness of endogenous punishment to change driving styles, our elicitation of subjects' beliefs allows us to connect the mechanism by which it works to influence via social norm formation. Indeed, according to Bicchieri ([3]) there is a social norm in place when people in a group expect others to conform to the behavior dictated by the norm, and in addition they expect others to enforce that behavior on them through punishment. As we discussed above, in the *Endogenous* condition subjects predict accurately the behavior of the group (thus, their empirical expectations are correct) while their beliefs are incorrect in the other conditions. In addition, in the *Endogenous* condition subjects know that they can be punished by their peers, instead of the external agent of the *Exogenous* condition, which turns more cautious behavior into a true social norm at least in a sizable fraction of the population. Therefore, our results suggest that a proper policy to prepare for mixed-agency scenarios is through behavior change interventions that appeal directly to people's expectations about what others will do ([8]). The *Endogenous* condition did not, however, yield higher social welfare, because the deterrence of the most risky behaviors comes hand in hand with a reduction in the least risky behaviors. This is a reminder that safety policies lead to behavioral adjustments that can yield unexpected consequences ([1]). Furthermore, suggests that regulators may need to promote only the best behaviors as being the acceptable social norm.

References

[1] Adams, John GU. "Seat belt legislation: the evidence revisited." Safety Science 18.2 (1994): 135-152.

[2] Andreoni, James. "Satisfaction Guaranteed: When Moral Hazard Meets Moral Preferences." American Economic Journal: Microeconomics *forthcoming*.

[3] Bicchieri, Cristina, *The Grammar of Society: The Nature and Dynamics of Social Norms* (Cambridge University Press, 2006).

[4] Bigoni, Maria, Gabriele Camera, and Marco Casari. "Partners or Strangers? Cooperation, monetary trade, and the choice of scale of interaction." American Economic Journal: Microeconomics *forthcoming*.

[5] Blincoe, Lawrence, Angela Seay, Eduard Zaloshnja, Ted Miller, Eduardo Romano, Stephen Luchter, and Rebecca Spicer. "The economic impact of motor vehicle crashes, 2000." DOT HS 809 (2002): 446.

[6] Borcan, Oana, Mikael Lindahl, and Andreea Mitrut. "Fighting corruption in education: What works and who benefits?" American Economic Journal: Economic Policy 9.1 (2017): 180-209.

[7] Camera, Gabriele, and Marco Casari. "The coordination value of monetary exchange: Experimental evidence." American Economic Journal: Microeconomics 6.1 (2014): 290-314.

[8] Christmas, Simon, Michie, Susan, and West, Robert, eds, *Thinking about behavior change: an interdisciplinary dialogue* (UCL Centre for Behavior Change, 2015).

[9] Chen, Yan, Fangwen Lu, and Jinan Zhang. "Social comparisons, status and driving behavior." Journal of Public Economics 155 (2017): 11-20.

[10] DeAngelo, Gregory, and Benjamin Hansen. "Life and death in the fast lane: Police enforcement and traffic fatalities." American Economic Journal: Economic Policy 6.2 (2014): 231-57.

[11] Egas, Martijn, and Arno Riedl. "The economics of altruistic punishment and the maintenance of cooperation." Proceedings of the Royal Society of London B: Biological Sciences 275.1637 (2008): 871-878.

[12] Fehr, Ernst, and Simon Gächter. "Cooperation and punishment in public goods experiments." American Economic Review 90.4 (2000): 980-994.

[13] Gächter, Simon, Elke Renner, and Martin Sefton. "The long-run benefits of punishment." Science 322.5907 (2008): 1510-1510.

[14] Green, Colin P., John S. Heywood, and Maria Navarro. "Traffic accidents and the London congestion charge." Journal of Public Economics 133 (2016): 11-22.

[15] Habyarimana, James, and William Jack. "Heckle and Chide: Results of a randomized road safety intervention in Kenya." Journal of Public Economics 95.11-12 (2011): 1438-1446.

[16] Holt, Charles A., and Susan K. Laury. "Risk aversion and incentive effects." American economic review 92.5 (2002): 1644-1655.

[17] Ito, Koichiro, Takanori Ida, and Makoto Tanaka. "Moral Suasion and Economic Incentives: Field Experimental Evidence from Energy Demand." American Economic Journal: Economic Policy 10.1 (2018): 240-67.

[18] Jakob, Michael, Dorothea Kbler, Jan Christoph Steckel, and Roel van Veldhuizen. "Clean up your own mess: An experimental study of moral responsibility and efficiency." Journal of Public Economics 155 (2017): 138-146.

[19] Masclet, David, Charles Noussair, Steven Tucker, and Marie-Claire Villeval. "Monetary and nonmonetary punishment in the voluntary contributions mechanism." American Economic Review 93, no. 1 (2003): 366-380.

[20] Nikiforakis, Nikos, and Hans-Theo Normann. "A comparative statics analysis of punishment in public-good experiments." Experimental Economics 11.4 (2008): 358-369.

[21] Noussair, Charles, and Steven Tucker. "Combining monetary and social sanctions to promote cooperation." Economic Inquiry 43.3 (2005): 649-660.

[22] Rege, Mari, and Kjetil Telle. "The impact of social approval and framing on cooperation in public good situations." Journal of Public Economics 88.7 (2004): 1625-1644.

[23] "Speed Management." Paris, France: Organisation for Economic Co-operation and Development; 2006. http://www.itf-oecd.org/sites/default/files/docs/06speed.pdf, accessed Oct 2, 2018.

[24] van Benthem, Arthur. "What is the optimal speed limit on freeways?" Journal of Public Economics 124 (2015): 44-62.

[25] Weng, Qian, and Fredrik Carlsson. "Cooperation in teams: The role of identity, punishment, and endowment distribution." Journal of Public Economics 126 (2015): 25-38.

[26] World Health Organization. Global status report on road safety 2015. World Health Organization, 2015.

[27] http://www.driverless-future.com/?page_id=384.

[28] "GM targets 2019 for U.S. launch of self-driving vehicles." https://business.financialpost.com/transportation/gm-targets-2019-for-u-s-launch-of-self-driving-vehicles, accessed on Oct 4, 2018.

[29] "Ford aims for self-driving car with no gas pedal, no steering wheel in 5 years, CEO says." https://www.cnbc.com/2017/01/09/ford-aims-for-self-driving-car-with-no-gas-pedal-no-steering-wheel-in-5-years-ceo-says.html, accessed on Oct 4, 2018.

[30] "Volvo Cars Plans a Self-Driving Auto by 2021." https://www.bloomberg.com/news/articles/2016-07-22/volvo-cars-plans-a-self-driving-auto-by-2021-challenging-bmw, accessed on Oct 4, 2018.

[31] "BMW says self-driving car to be Level 5 capable by 2021." http://www.autonews.com/article/20170316/MOBILITY/170319877/bmw-says-self-driving-car-to-be-level-5-capable-by-2021, accessed on Oct 4, 2018.

Appendix A

Proof of proposition 1

$$S_F = 2, S_S = 1, S_A = 0.5; a_F = 0.35, a_S = 0.3, a_A = 0$$

$$\begin{aligned} E(U(F)) &= 2^{\gamma}\left(1 - 0.35\left(2x_F + x_S + 0.5\left(1 - x_F - x_S\right)\right)\right) \\ E(U(S)) &= \left(1 - 0.3\left(2x_F + x_S + 0.5\left(1 - x_F - x_A\right)\right)\right) \\ E(U(A)) &= 0.5^{\gamma} \end{aligned}$$

For part 1 of the proposition to be true we need to show that action S is not a best response for all possible beliefs about the population ($AS \in [0.5, 2]$) and for reasonable risk preferences ($\gamma \in (0, 1)$). Suppose not, then there is some γ and AS for which

$$E(U(S)) > \max\left\{E(U(A)), E(U(F))\right\}.$$

$$\begin{aligned} 2^{\gamma}\left(1 - 0.35AS\right) &< 1 - 0.3AS \\ 0.5^{\gamma} &= \frac{1}{2^{\gamma}} < \left(1 - 0.3AS\right) \Leftrightarrow \frac{1}{\left(1 - 0.3AS\right)} < 2^{\gamma} \\ \frac{\left(1 - 0.35AS\right)}{\left(1 - 0.3AS\right)} &< 2^{\gamma}\left(1 - 0.35AS\right) < 1 - 0.3AS \end{aligned}$$

For such values to exist, it is necessary that

$$1 < \frac{\left(1 - 0.3AS\right)^2}{\left(1 - 0.35AS\right)}$$

The derivative of $\frac{\left(1-0.3AS\right)^2}{\left(1-0.35AS\right)}$ with respect to AS is

$$\frac{-0.6\left(1 - 0.3AS\right) + 0.35\left(1 - 0.3AS\right)}{\left(1 - 0.35AS\right)^2} = \frac{0.075\,AS - 0.25}{\left(1 - 0.35AS\right)^2} < 0$$

and thus $\frac{\left(1-0.3AS\right)^2}{\left(1-0.35AS\right)} < 1$ for $AS \in [0.5, 2]$.

For part 2, notice that if an entire population chooses F, then it must be the case that for all $\gamma \in (0,1)$

$$\begin{aligned} 2^{\gamma}(1-0.35\cdot 2) &= 0.3\cdot 2^{\gamma} \geq 0.5^{\gamma} \\ 0.3 &\geq 0.25^{\gamma}. \end{aligned} \tag{2}$$

Since inequality 2 is only true for $\gamma \geq \frac{-\ln 0.3}{-\ln 0.25} \simeq 0.868\,48$, then there is a contradiction.

If an entire population chooses A, then it must be the case that for all $\gamma \in (0,1)$

$$\begin{aligned} 2^{\gamma}(1-0.35\cdot 0.5) &= 0.825\cdot 2^{\gamma} \leq 0.5^{\gamma} \\ 0.825 &\leq 0.25^{\gamma}. \end{aligned} \tag{3}$$

Again, since 3 is only true for $\gamma \leq \frac{-\ln 0.825}{-\ln 0.25} \simeq 0.138\,77$, we have a contradiction. □

Proof of Proposition 2

Denote AS_i^P as the equilibrium belief of driver i about the average speed of the population under the punishment system P,p. Similarly, denote AS_i as the belief of driver i about the average speed of the population without the system P,p. Proposition 1 states that the equilibrium will consist of a mixture of F and A drivers (S will never be chosen). The proportions of drivers choosing actions F and A are determined by the set of drivers that strictly prefer one action over the other. Since both sets have positive measure, and all $\gamma_i \in (0,1)$ have positive measure, there will be a type i driver who is indifferent between the two actions. For that driver, in a system without punishment, it must be that case that

$$U_i(S_F)(1-a_F AS_i) = U_i(S_A)(1-a_A AS_i).$$

With the punishment according to the P,p system, it must be that case that

$$((1-p)U_i(S_F) + pU_i(S_F-P))\left(1-a_F AS_i^P\right) = U_i(S_A)\left(1-a_A AS_i^P\right).$$

In a system without punishment, a driver will chose action F if

$$\frac{U_i(S_F)}{U_i(S_A)} = \frac{(1-a_A AS_i)}{(1-a_F AS_i)}. \tag{4}$$

In a system with punishment, a driver will choose action F if

$$\frac{((1-p)U_i(S_F) + pU_i((S_F-P)))}{U_i(S_A)} = \frac{\left(1-a_A AS_i^P\right)}{(1-a_F AS_i^P)}. \tag{5}$$

When comparing the left side of equations 4 and 5, it is clear that

$$\frac{U_i\left(S_F\right)}{U_i\left(S_A\right)} > \frac{\left(\left(1-p\right)U_i\left(S_F\right)+pU_i\left(\left(S_F-P\right)\right)\right)}{U_i\left(S_A\right)}.$$

This means that the following must be true:

$$\frac{\left(1-a_A AS_i\right)}{\left(1-a_F AS_i\right)} > \frac{\left(1-a_A AS_i^P\right)}{\left(1-a_F AS_i^P\right)}$$

This immediately implies the following comparison between the Average Speeds across the two systems.

$$\begin{aligned}
\left(1-a_F AS_i^P\right)\left(1-a_A AS_i\right) &> \left(1-a_A AS_i^P\right)\left(1-a_F AS_i\right) \\
-a_F AS_i^P - a_A AS_i &> -a_A AS_i^P - a_F AS_i \\
\left(a_F-a_A\right)AS_i &> \left(a_F-a_A\right)AS_i^P \\
AS_i &> AS_i^P
\end{aligned}$$

Note that $AS = x_F S_F + x_S S_S + x_A S_A$. Proposition 1 states that x_S is zero without $P, p,$, which means that it must be the case that for $AS_i > AS_i^P$ to be true, we must have that $x_F^P < x_F$.□

Appendix B - Subject statistics

Table 6: Subject statistics by policy condition

Treatment	# of subjects	Avg. time of day	Avg. # of rounds	# of choices	% male	% undergrad	Avg. age
Control	80	12:48	21.6	1,735	49%	54%	24.2
Framing	77	12:15	18.9	1,457	53%	40%	24.4
Exogenous	84	12:33	20.4	1,703	43%	46%	24.4
Endogenous	85	12:48	21.8	1,854	46%	42%	23.8
Total	326	12:36	20.7	6,749	48%	46%	24.2

Appendix C - *Endogenous* punishment analysis

Our main results suggest that the availability of endogenous punishment was the only policy that significantly alters driving behavior. Because of this, the *Endogenous* condition deserves a deeper analysis. *What type of subjects are contributing to the (costly) punishment fund?* Out of a possible 1,854 contribution choices, subjects contributed to punishment on 209 occasions (11.3%). Table 7 shows the punishment decisions by driving choice. This table shows that Slow drivers were most likely to contribute to the punishment fund, however this result is only marginally significant (Fisher's Exact Test p-val <0.1 for each pair-wise comparison with Auto and Fast).

Table 7: Punishment contributions by driving choice

Punishment contributions				
Contribute	Auto	Slow	Fast	Total
Yes	62	61	86	209
No	528	378	739	1,645
Perc	10.5%	13.9%	10.4%	11.3%

Table 8 classifies subjects based on the number of times they contributed to the punishment fund. Approximately half of the subjects contributed to the fund at least once whereas three subjects contributed in every round they played. Interestingly, the three "Always" subjects employed the same driving choice in every round they played. One subject chose Fast every round, one subject chose Slow every round, and one subject chose Auto every round. All three "Always" subjects were female graduate students.

Table 8: Number of punishment contributions per subject

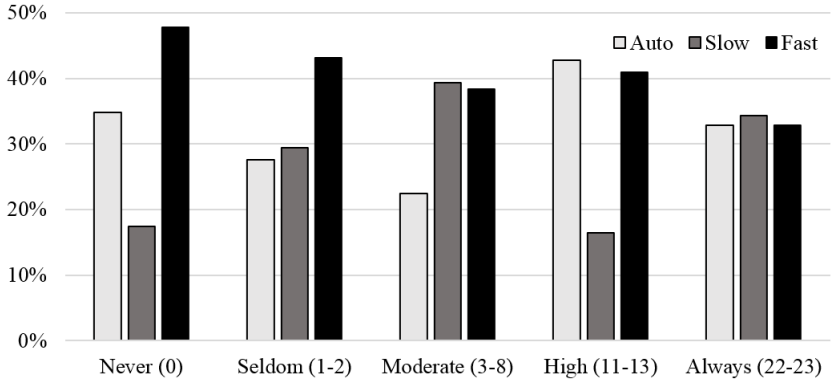
	Subject punishment types				
Contribution type	Never 0	Seldom 1-2	Moderate 3-8	High 11-13	Always 22-23
# of subjects	44 (51.7%)	23 (27.1%)	10 (11.8%)	5 (5.9%)	3 (3.5%)
% of total contributions	0	16.3%	23%	28.7%	32.1%

Table 9 and Figure 1 investigate the driving choices within the punishment types described in Table 8. There doesn't seem to be a monotonic relationship between the number of punishment contributions and any of the three driving choices. From this analysis, it seems that punishing subjects are not substantially different from non-punishers in their driving choices. The decision to endogenously punish is orthogonal to the choice about driving.

Table 9: Number of punishment contributions per subject

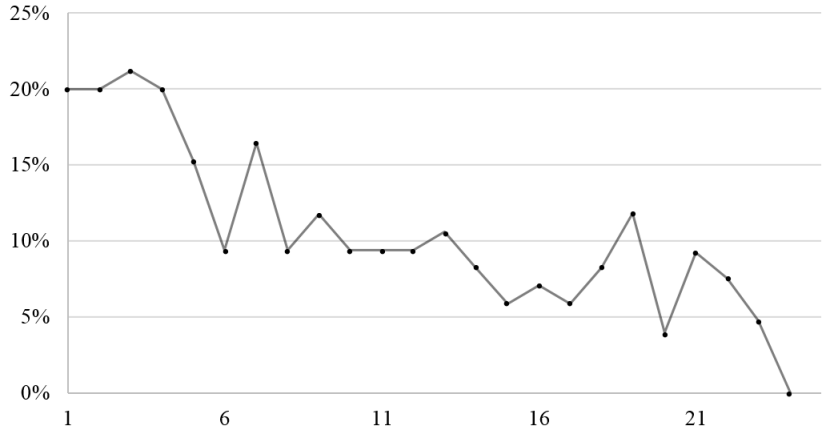
	Subject choices by punishment types				
Contribution type	Never 0	Seldom 1-2	Moderate 3-8	High 11-13	Always 22-23
Auto	336 (34.8%)	137 (27.6%)	48 (22.4%)	47 (42.7%)	22 (32.8%)
Slow	168 (17.4%)	146 (29.4%)	84 (39.3%)	18 (16.4%)	23 (34.3%)
Fast	462 (47.8%)	214 (43.1%)	82 (38.3%)	45 (40.9%)	22 (32.8%)
# of choices	966	497	214	110	67

Figure 1: Subject choices by punishment types



As is expected in public goods type games, costly punishment declines with time. Figure 2 shows that the proportion of subjects contributing to the punishment fund declines over the duration of the experiment.

Figure 2: Contribution decisions by round



Appendix D - Screen shots

Figure 3: *Control* choice screen

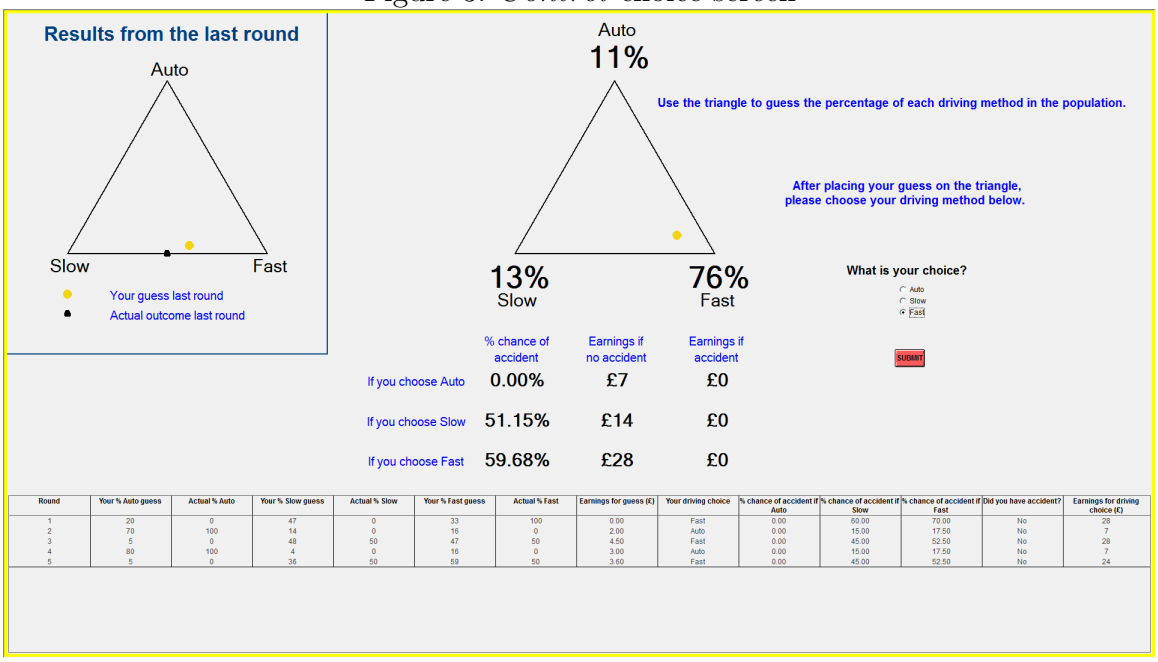


Figure 4: *Framing* choice screen

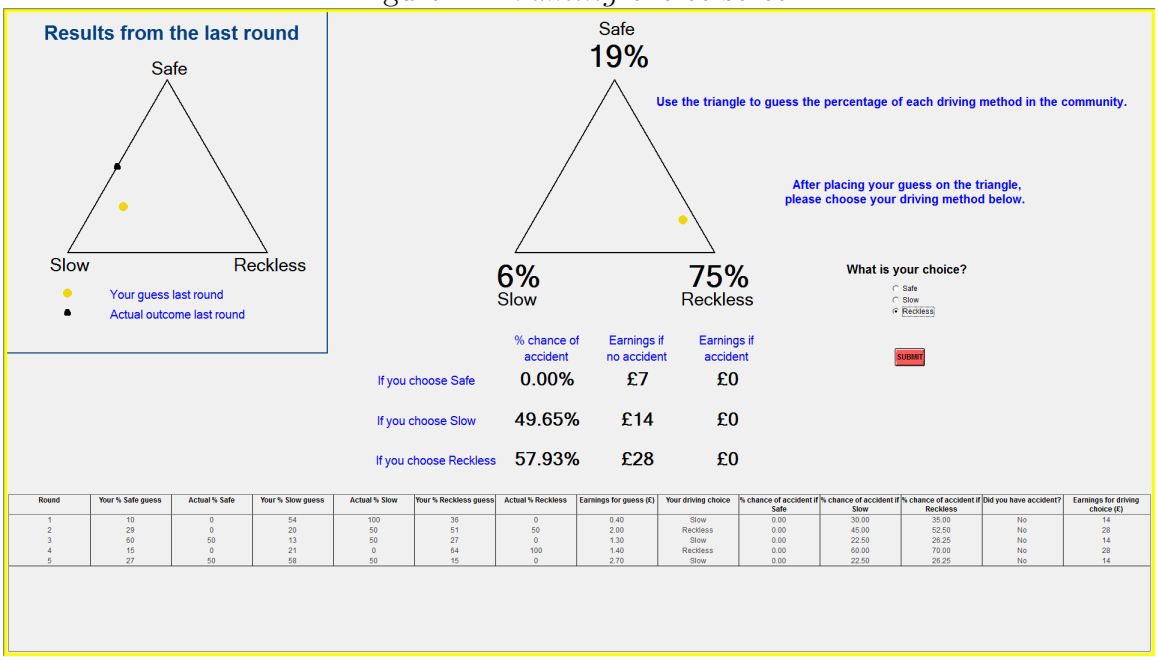


Figure 5: *Exogenous* punishment choice screen

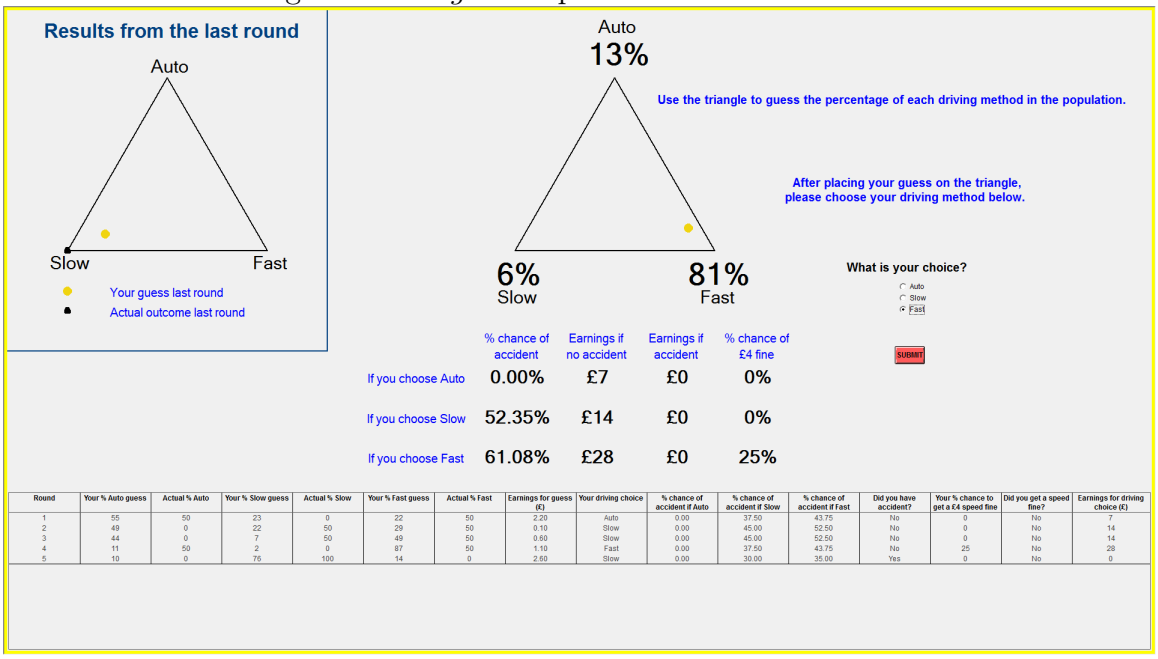


Figure 6: *Endogenous* punishment choice screen

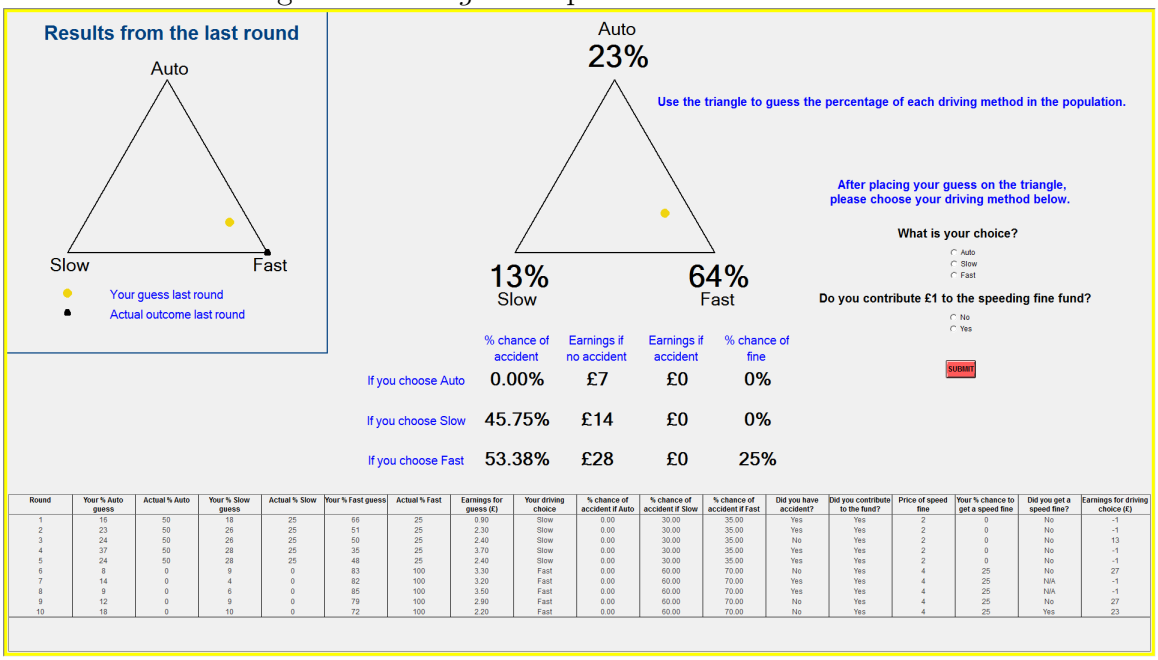


Figure 7: *Control* results screen

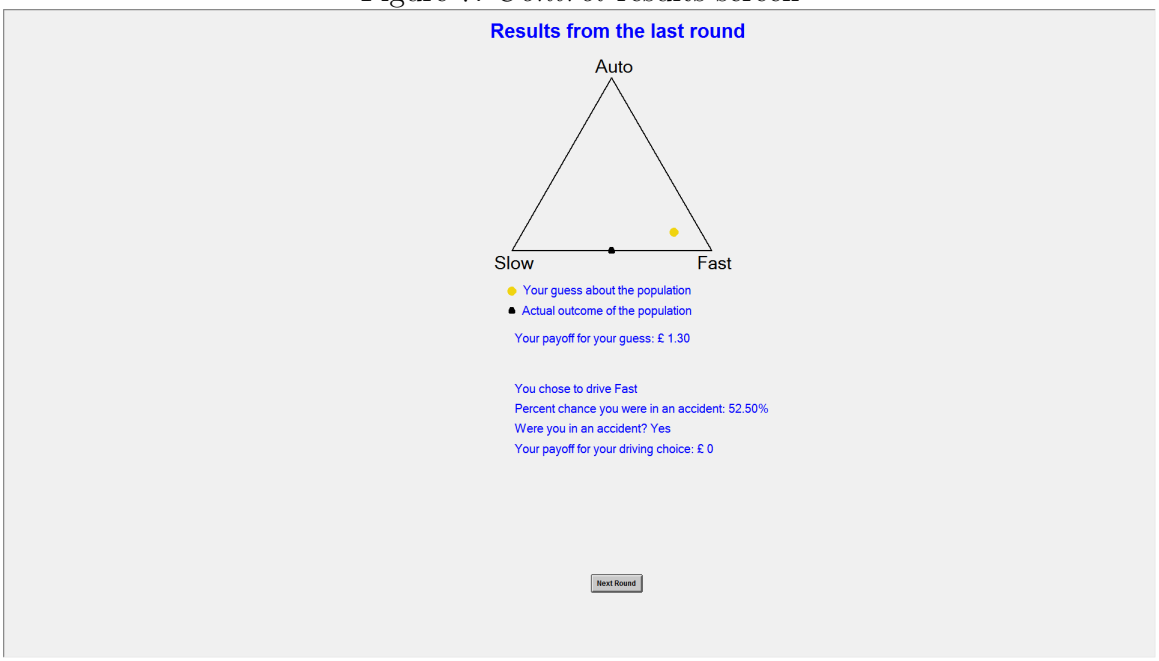


Figure 8: *Framing* results screen

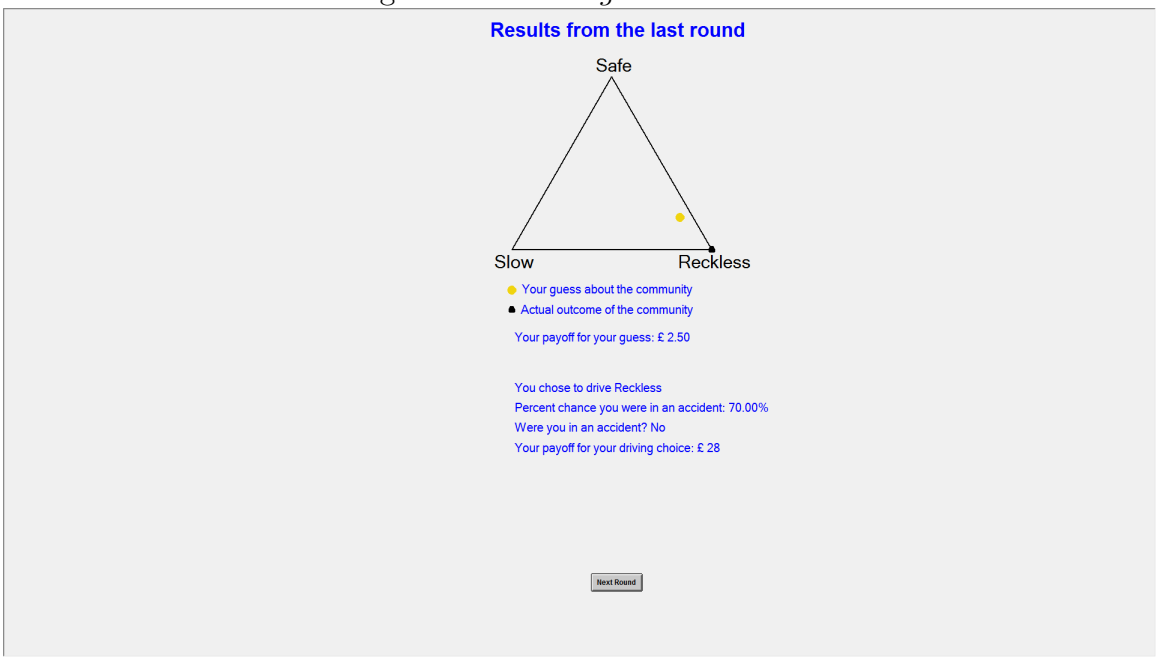


Figure 9: *Exogenous* punishment results screen

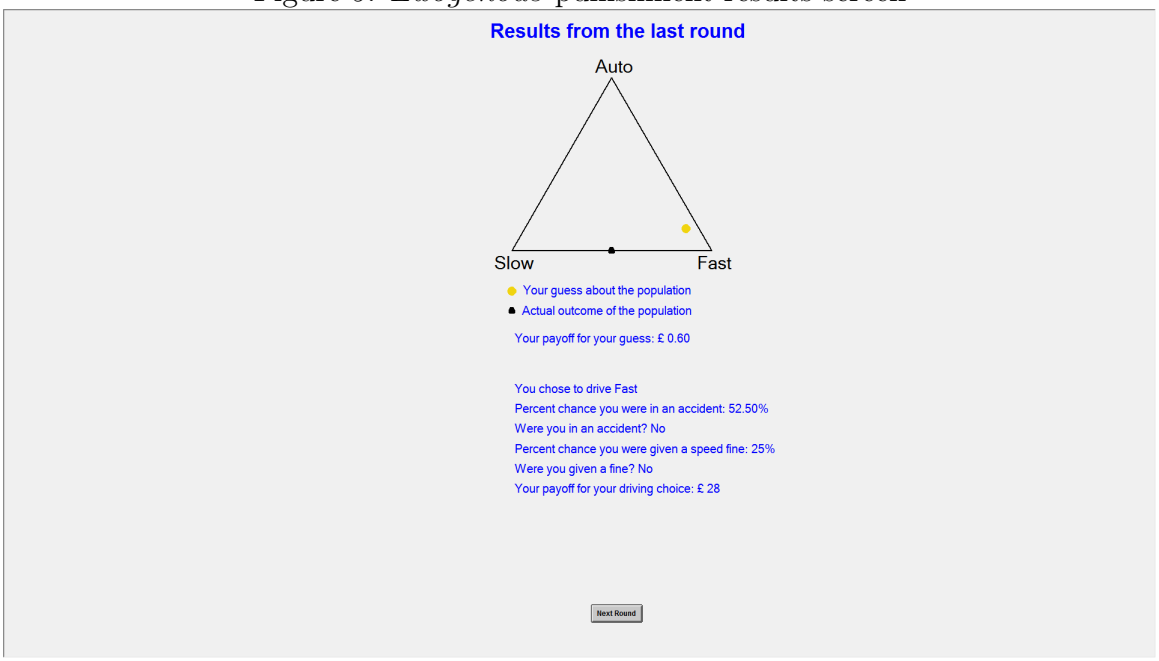


Figure 10: *Endogenous* punishment results screen