Gender differences in the effectiveness of prosocial policies: an application to road safety

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

Working document as of April 1, 2020

Abstract

We study policies aimed at discouraging behavior that produces negative externalities, and their differential gender impact. Using driving as an application, we develop a model where slowest vehicles are the safest choice, whereas faster driving speeds lead to higher potential payoffs but higher probabilities of accidents. Faster speeds have a personal benefit, but create a negative externality. The model motivates four experimental policy conditions. We find that the most effective policies use different framing and endogenously determined punishment mechanisms (to fast drivers by other drivers). These policies are only effective for female drivers which leads to substantial gender payoff differences.

[*]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

In this paper, we use a theoretical and experimental approach to analyze the effectiveness of policies aimed at discouraging antisocial behavior in mixed-agency environments. We focus on a specific application in which a negative externality is imposed upon a population by individuals who choose to drive fast. As described in more detail in Section 2, previous literature suggests that such policies could have differential gender effects, but their size and direction are not entirely clear. Thus, our main question of interest is to measure the differential effect on genders of policies designed to promote cooperative human behavior in mixed-agency driving environments.

Certain individual behaviors create detrimental negative impacts on a society. One such behavior, which motivates our study, is unsafe driving habits. Excessive speed is the number one road safety problem in most countries ([35]), and men, especially young men, are disproportionately involved in accidents ([33]). A recent field experiment quantifies the relationship between speed and negative outcomes ([38]). 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. 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). While everyone will use different driving speeds, these choices are interrelated. One driver's excessive speed creates a more dangerous driving situation for everyone involved.

The issues related to driving styles and their related externalities are about to become more pressing in the near future. The reason is that automation will soon drastically change transportation. While pioneered by Tesla, even mass producing car companies such as BMW([43]), Ford ([41]), GM ([40]), and Volvo ([42]) expect fully automated models on the road by 2021. Autonomous vehicles are inherently safer, but their safety might encourage free-riding among other drivers. Thus, given this impending change, policies encouraging safe driving in this "mixed-agency" environment are particularly time-sensitive.

In Section 3, we develop a game theoretic model of a driving scenario where drivers choose between two manual driving styles ("Fast" or "Slow") and one style which allows their vehicle to drive automatically ("Auto"). We assume that Auto drivers are never in an accident and will therefore earn a constant amount. However, for each individual driver, faster (manual) speeds lead to higher potential payoffs with higher probabilities of being in an accident. With risk-neutral (or slightly risk-averse) drivers, the fastest driving speed (Fast) is a dominant strategy. However, faster driving speeds also increase the probability that *all*

drivers are involved in an accident, thus creating a negative externality on the population. Policies can play a role in discouraging individual drivers from free-riding off of the safety provided by others' safer driving styles.

As described in Section 4, our model is used to parametrize the control condition of a laboratory experiment ("*Control*"). The observed behavior in *Control* 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 ([34]), one condition uses associative language to encourage cooperative driving behavior ("*Framing*"). In addition, we conduct two treatment conditions using punishment: "*Exogenous*" and "*Endogenous*". *Exogenous* is the same as *Control* with the addition that participants choosing Fast have the possibility of incurring an exogenously determined financial penalty. *Endogenous* includes a similar probabilistic fine, but the fine amount is determined by contributions made by other drivers.

Section 5 presents the results. Our main experimental finding is that no policy has an effect on male participants in terms of reducing the most dangerous driving style, Fast. However, all policy conditions have the same effect on female participants - they choose Fast less often and Auto more often. This effect is particularly salient in *Endogenous* and *Framing*.

Our experimental design allows us to probe into different possible mechanisms to explain these findings. In our experiment, participants submitted (incentivized) beliefs about the proportion of driving choices selected by other participants. One possibility is that the effect is mediated by these beliefs. In fact, we observe that overall beliefs become more accurate in any policy condition. However, both genders are equally accurate at predicting the driving choices of others in all conditions. Therefore, differences in the accuracy of beliefs cannot explain differences in driving behavior. However, while beliefs are in general equally (in)accurate, there is a difference across genders. We find that in any policy condition, female participants believe Auto will be chosen more often, particularly so in *Endogenous* (and marginally so in $Framing$). When we combine this information with our questions pertaining to first-order and second-order norms, we arrive at a likely explanation for our results, namely that in the presence of any policy, the social norm of female participants is more likely to be Auto than the social norm of male participants. Similarly, in the presence of any policy, the social norm of male participants is more likely to be Fast than the norm of female participants. This points to social norms as the reason underlying the differences in choices. To further support this, we show that female participants are more likely to switch to non-Fast driving choices after receiving a driving fine. This suggests that female participants understand that choosing Fast deviates from the social norm, which deserves

punishment, which reinforces our interpretation. In addition, the fact that *Endogenous* has the largest effect on behavior and beliefs also supports our narrative focused on social norms. Many previous papers show that social norms are more deeply incorporated when they are formed endogenously ([2], [7], [9], [21], [24], [28], [32]).[1]

Interestingly, the average total payoffs do not change in *Endogenous* or *Framing*. Our data suggest the following explanation. *Endogenous* and *Framing* show lower numbers of drivers choosing faster styles (most of them female participants) which, in turn, incentivize other drivers to switch into faster (manual) driving choices (most of them male participants). This result is consistent with evidence that some vehicle safety measures (such as seat belts or ABS brakes) do not save lives because the introduction of the extra security corresponds with drivers choosing more aggressive driving styles ([1]). We believe this information is relevant for policy makers. The policies we establish to combat behaviors leading to negative externalities may be ineffective and, in addition, the behavioral reactions in the population could increase gender inequity. Since we uncover social norms as a mediating mechanism, this suggests that interventions specially targeted at male participants' socially dysfunctional norms could be particularly promising.

2 Relevant literature

In the setting we study, it is possible that governmental regulation or community enforcement may be helpful in promoting cooperation in terms of safe driving choices. As we see it, our study contributes to work in two areas: problems of enforcement of prosocial behavior, and the effectiveness of prosocial policies in real-world (driving) contexts. In both of them there is an important (but often ambiguous) gender effect that we also discuss.

Our focus is related to a long tradition of studying human cooperation. Unsurprisingly, punishment is highly effective in enforcing social norms ([19], [20], [30], [32], and [31]). A recent field experiment shows that external punishment (along with monitoring) can decrease bribing behavior in education ([8]). In addition, moral suasion is also a powerful mechanism to develop social norms around prosocial behavior. Another field experiment testing the policy effectiveness in the domain of energy demand shows that the combination of moral suasion and economic incentives produce substantially different policy impacts ([27]).

There are interesting, but inconsistent, differences along gender lines relating to prosocial behavior. Female participants are more averse to inequality ([14]) and less likely to lie or

[1] In agreement with the social norm being stronger for female participants in our study, they contribute to the (costly) punishment of Fast drivers at a slightly higher rate than male participants (13.5% versus 8.6%). However, this difference is not statistically significant in a logistic regression. Further analysis on the punishment decisions in *Endogenous* are in Appendix E.

cheat for monetary benefit ([16] and [23]). Male participants are more likely to violate the social norm when they can do so privately ([29]). In addition, evidence has suggested that the neural correlates for social norm compliance are systematically different across genders ([10]). In spite of this, when analyzing situations closest to our study, there is plenty of mixed evidence pertaining to the level of prosociality between genders. For example, a review of public goods experiments show that gender differences are not straightforward and that the context plays a crucial role ([17]). A similar nuanced story is shown for punishment to free riders ([18]) and for charitable donations ([4]). Female participants are more prone to donate in dictator games when it is more costly to themselves, whereas male participants donate more when it is cheap ([3]). Finally, whether it is true or not, participants expect female participants to be more altruistic than male participants, which implies a connection between expected behavior and compliance with social norms ([5]). Because this literature provides ambiguous messages about the effect of gender, we create a model in Section 3 that does not directly account for gender differences. Instead we allow the data to clarify the direction of effects (Section 5).

Policy levers that discourage unsafe driving behavior can have immense societal benefits. A central concern for people focused on driving safety is to understand what type of incentive is effective in this setting. Unsurprisingly, there is a close relationship between prosocial 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 ([15]). Fines can be particularly effective to deter traffic violations by women ([39]). Endogenous 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 ([12]). Endogenous intra-group pressure can be particularly effective in enforcing a social norm. For example, a 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 ([24]). In addition, previous studies have show that "males, on average, felt less confident in their ability to influence other drivers and perceived more costs if doing so than females did." ([37]).

We find that one of our most effective policies uses an endogenous mechanism to enforce prosocial behavior (*Endogenous*). This aligns with previous findings in other problems of strategic uncertainty. For example, *Endogenous* combines monetary and social sanctions, which are typically salient in promoting cooperation ([32]). In addition, this result is 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 ([24]).

Giving participants the agency in the punishment process shifts the moral responsibility to solve the problem endogenously. In addition, allowing monetary exchange systems to endogenously emerge can support a social norm of cooperation in large groups ([7] and [9]). Finally, previous research suggests that people are responsive to their "moral responsibility" in settings where each others' actions affect the population ([28]). This paper also shows that female participants can take their responsibility more seriously ([28]). On the other hand, as mentioned above, female participants are shown to be less likely to punish behavior inconsistent with the social norm ([18]). We aim to clarify which effects dominate in our setting.

3 Theory

We model the availability of three different driving styles, with different safety levels for the collective group of drivers. The safer styles, take actions whenever they risk colliding with another vehicle. This added safety comes at the cost of personal speed. On a road with safer vehicles, less safe drivers may free-ride off of the fact that others will prioritize safety over speed.

We aim to study which policies can mitigate the free-riding problem in this environment. 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, for example, autonomous vehicles which are programmed to put safety as a top concern.

3.1 The game

We denote by S_i the average speed of an agent choosing driving style $i \in \{F, S, A\}$ (F stands for Fast, S stands for Slow, and A for Automated). We assume that the driving speed 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 an accident.

$$p_i = a_i AS$$

We assume that the probability of an accident 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.

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

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

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

The above model depends on the following parameters: the average speeds (S_F, S_S, and S_A), the accident 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 accident probabilities that will be implemented in the experiment. 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 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 participants 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.*

3.2 Theoretical implications of policy conditions

The game and hypotheses derived in the previous section will serve as our control condition of the experiment ("*Control*"). The main interest of the paper is to test the effectiveness of different policy conditions 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 ([15] and [22]). 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 ([34]) as well as in real-world driving environments ([12] and [24]). 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

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

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). As with *Exogenous*, the government imposes a (probabilistic) penalty for choosing action F. However, in this condition, drivers can, at a personal cost, 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 ([7], [9], [28], and [32]). 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.*

Hypotheses 3, 4, and 5 predict that the proportion of F will be lower in each respective policy condition. While outside of our formal model, it may be true that different sub-groups respond differently to these policy conditions. We would expect this to be particularly true in *Endogenous* and *Framing*, as they involve violations of social norms. Female participants have been shown to be relatively more sensitive in responding to these types of social cues in a variety of contexts similar to ours [14].

4 Experimental design

4.1 Participants and sessions

Experiments are conducted at a large public university. Each participant interacts in one policy condition. We conduct 8 sessions for each condition for a total of 32 experimental

sessions. Each session consists of between 8 and 12 participants and lasts no longer than 2 hours. At the end of each session participants provide demographic information about gender, risk preference, age, and experience with driving. Appendix B further describes the data for all 326 participants and checks to ensure that our conditions are balanced across the demographic variables.

4.2 Task

After instructions and a test of comprehension, participants interact in a multi-round decision-task. In order to avoid strange behavior associated with the final round of the session, the number of rounds is randomly determined to be between 17 and 25 and the participants do not know which round will be the final one in their session.[2] In each round, participants make 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 F.

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

In each round, every participant submits a guess about the proportion of participants in the room who will choose Fast, Slow, and Auto. They do so by using the computerized "triangle tool" which allowed participants to make their guess by dragging a point within a triangle where each vertex of the triangle represents a guess where 100% of the participants in the room are choosing one driving style. The amount a participant earns from their guess is £5 minus the difference between their guessed distribution of driving styles and the actual distribution of driving styles in that round. A perfect guess earns £5 and a very inaccurate guess earns £0.

The triangle tool is also used by participants to calculate the probability of an accident for each driving style conditional on a possible distribution of driving styles. The proba-

[2]Starting in round 18, there is a $\frac{2}{3}$ chance that another round will be played. This process continues until round 25 is reached, which is determined to be the last round. Participants are 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.

bility of being in an accident (and earning £0) for each driving style is updated when the participant changes their guess about the population. This way, participants can 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, participants have complete information about their choices, the choices of other participants in the room, and their payoffs in all previous rounds. In addition, a picture is shown in the top-left of the screen which shows the distribution of driving style choices in the previous round as well as that participant's guess about the distribution in the previous round.

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

4.3 Policy conditions

Control. Participants interact in the experiment described above. Participants choose between driving "Fast", "Slow", or "Auto" and are incentivized to guess the distribution of these driving types within the "population" of other participants.

Framing. Participants choose between driving "Reckless", "Slow", or "Safe" and are incentivized to guess the distribution of these driving types within the "community" of other participants. This type of associative framing can increase contribution rates in public goods games ([34]).

Exogenous (punishment). Participants who choose Fast have a 25% chance to pay a fine of £4. This fine only applies to participants who are not in an accident in that round.

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

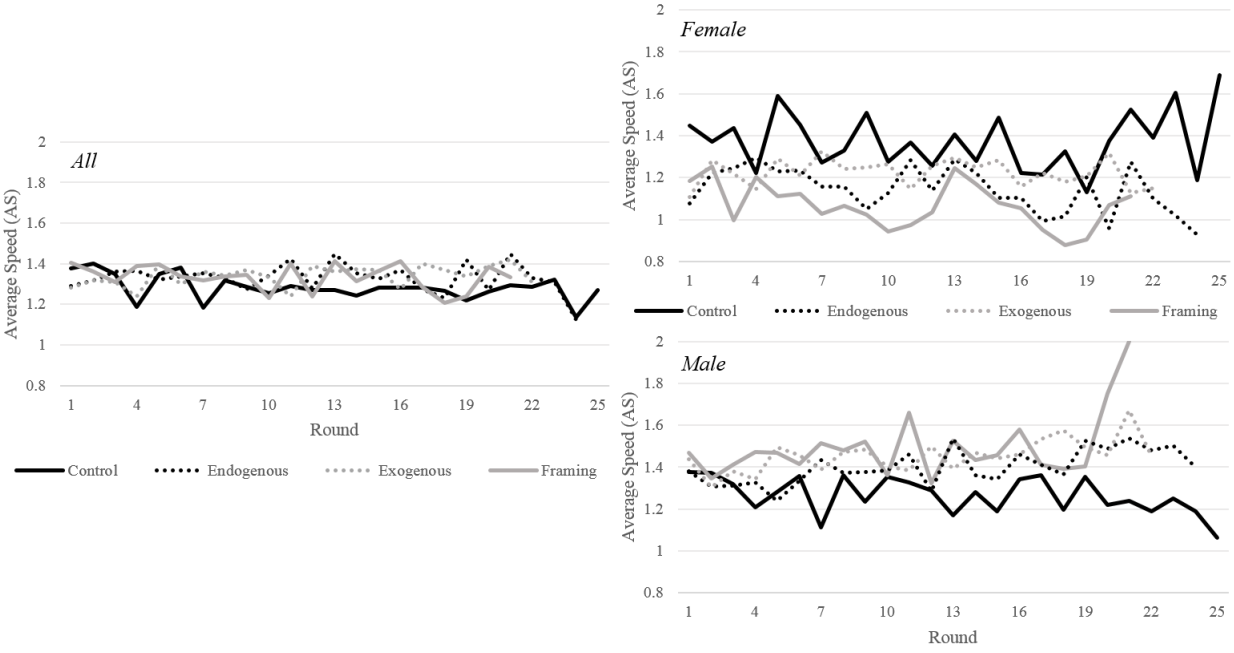
5 Results

5.1 Average Speed

In a session, the individual driving choices (of Fast, Slow, or Auto) determine the population's Average Speed (AS). AS is an important measure because it determines the probability of an accident for the Fast and Slow drivers. In this way, AS is a general measure of the overall safety of a driving environment.

Figure 1 plots the AS separated by condition.[3]

Figure 1: Average Speed by condition (left) and separated by gender (right)

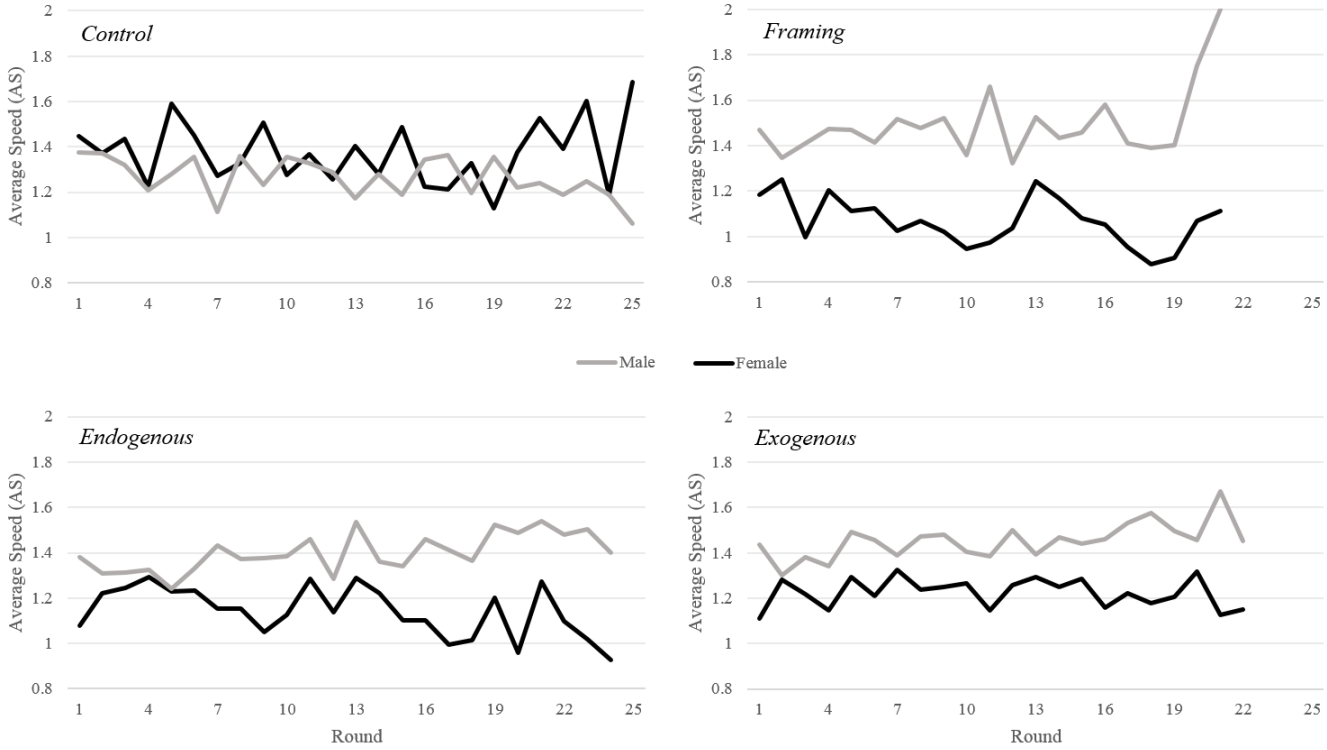


When analyzing all participants (left panel), it is clear that none of the policy conditions produce systematically lower AS than the AS observed in *Control*. However, all 3 policies have a striking effect when analyzing male and female participants separately (right panel). For female participants, the AS in all 3 conditions are lower than *Control* while for male participants, the AS in all 3 conditions are higher than *Control*.

Figure 2 plots the AS within each condition and separated by gender.

[3]AS=2 when all participants choose Fast and AS=0.5 when all participants choose Auto. Each line represents data from 8 sessions. Each point on a line is the average of all 8 AS observations in that round. For example, the 8 sessions in *Control* have AS observations in round 5 of 1.31, 0.88, 1.38, 1.31, 1.50, 1.20, 1.63, and 1.59. The average (1.35) is reported on the solid black line in round 5 in the left panel. Similar calculations can be made solely focusing on male or female participants in each session, which serves as the data for the right panel.

Figure 2: Average Speed within condition separated by gender



While 18 (out of 25) rounds in *Control* (top-left panel) report female participants with higher AS than male participants, neither gender chooses systematically higher AS. In the policy conditions, a strikingly different pattern emerges. For any round within any of the 3 policy conditions, it is always the case that the average AS of male participants is higher than the average AS of female participants.

To further explore this finding, we calculate the AS within each session averaged across all rounds (this yields one number for each session providing 8 numbers in a condition). In addition, we analyze the AS realizations pooling data from all 3 policy conditions ("AnyPolicy"; which has 24 realizations). Table 1 shows the average of these AS realizations separated by condition and gender along with p-values from two-sample t-tests.

Table 1: Average Speed by condition and gender

	Control	*Endogenous*	*Exogenous*	*Framing*	AnyPolicy
All participants	1.29	1.34	1.34	1.33	1.34
Male participants	1.28	1.40	1.44	1.45	1.43
Female participants	1.35	1.15	1.23	1.07	1.15
Male - Female	-0.07	0.25	0.22	0.38	0.28
Diff (p-value)	.413	.006	.090	.018	$< .001$

In *Control*, there is no difference in the AS across gender. In each policy condition, the AS of female participants is significantly lower than the AS of male participants ($p< 0.100$ for all 3 pairwise comparisons). This result is particularly strong in *Endogenous* and *Framing* ($p= 0.006$ and $p=0.018$, respectively). The strongest statistical significance is reached when pooling the data from all 3 policy conditions together which shows that female participants produce lower AS than male participants ($p< 0.001$).

Result (1)

In *Control*, the AS does not differ by gender. In each policy condition, the AS of female participants is lower than the AS of male participants. This is particularly salient in *Endogenous* and *Framing*.

5.2 Driving choices

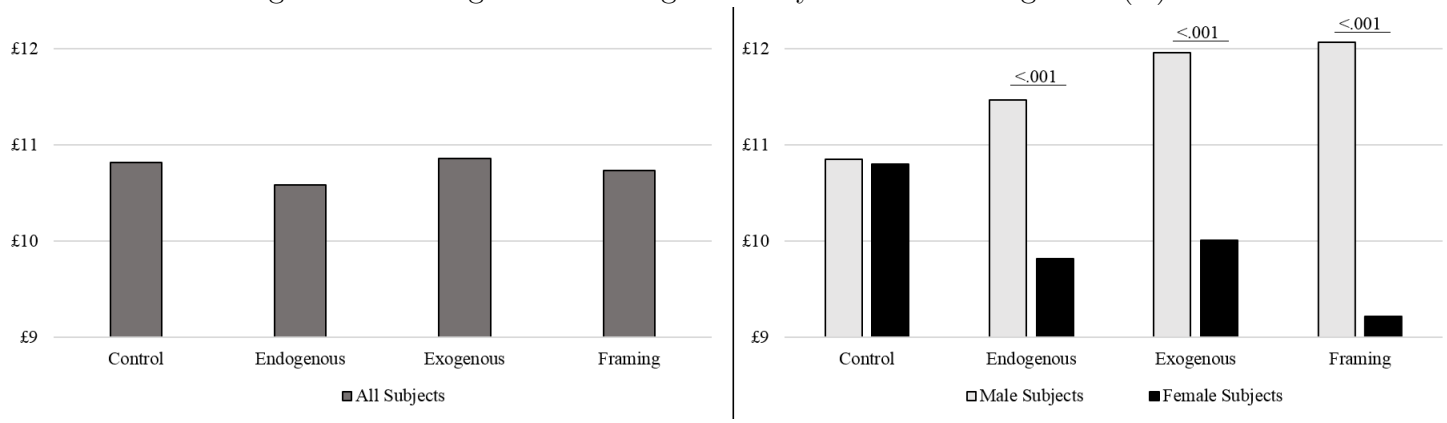
In each round, a participant makes a driving choice of either Fast, Slow, or Auto. Their driving choice and the driving choice of others will determine their earnings in that round.[4] The average earnings from these choices separated by condition and gender are shown in Table 2 and Figure 3.

Table 2: Earnings from driving choice by condition and gender (£)

	Control	*Endogenous*	*Exogenous*	*Framing*
All participants	10.82	10.58	10.86	10.73
Male participants	10.85	11.47	11.96	12.07
Female participants	10.80	9.82	10.01	9.22
Male - Female	0.05	1.65	1.95	2.85
Diff (p-value)	.918	$< .001$	$< .001$	$< .001$

[4]Participants also earn money for accurate beliefs about others, which is analyzed in the next section.

Figure 3: Earnings from driving choice by condition and gender (£)



When analyzing all participants (top row of Table 2 and left panel of Figure 3), none of the policy conditions produce significantly different average earnings relative to *Control* ($p > 0.480$ using a two-sample t-test for all 3 pairwise comparisons). Furthermore, earnings from driving choices in *Control* are not different across gender ($p = 0.918$). However, as with AS, separating by gender demonstrates a consistently different pattern, which yields our next result.

Result (2)

In *Control*, the earnings from driving choices do not differ by gender. In each policy condition, earnings from driving choices are lower for female participants when compared to male participants.

Results 1 and 2 demonstrate that policies have different effects on the AS and earnings of male and female participants. We now focus on driving choices, specifically, to further disentangle the effect of *Endogenous*, *Exogenous*, and *Framing*. Table 3 shows the percentage of driving choices observed across all rounds separated by condition and gender.

Table 3: Driving choices by condition and gender

	All			*Female*			*Male*		
Condition	% Fast	% Slow	% Auto	% Fast	% Slow	% Auto	% Fast	% Slow	% Auto
Control	46.7	19.7	33.6	49.0	21.8	29.1	44.3	17.5	38.2
Framing	46.0	21.0	33.0	33.8	23.3	42.9	56.9	19.0	24.2
Exogenous	48.7	20.8	30.5	41.4	22.3	36.3	58.2	18.8	23.0
Endogenous	44.5	23.7	31.8	34.7	30.8	34.6	56.2	15.2	28.6
AnyPolicy	46.4	21.9	31.7	36.9	25.8	37.4	57.1	17.8	25.4

When analyzing all participants (left panel), the profile of driving choices in *Framing* and *Exogenous* are not significantly different from *Control* (Pearson's Chi-Squared p = 0.665 and 0.144, respectively). *Endogenous* shows the largest effect with a profile of driving choices that is significantly different from *Control* at the 0.016 level. The pooled AnyPolicy condition is also significantly different from *Control* at the 0.001 level.

When analyzing all participants, the impact of each policy is (at best) rather small. However, as with AS (Result 1) and earnings from driving choices (Result 2), each condition has a large effect on driving choices within gender. The center panel of Table 3 shows that, compared to *Control*, each policy has less female participants who choose Fast and more female participants who choose Auto. This systematic effect go in opposite direction for male participants. As shown in the right panel of Table 3, compared to *Control*, each policy has more male participants who choose Fast and less male participants who choose Auto.[5]

We can further explore this relationship while controlling for independent variables. We have independent dummy variables for each conditions as well as a pooled AnyPolicy variable. In addition, after all of the driving choice rounds, participants are asked their gender ("Female"=1 if the participant self-reports as female) and are tasked with making incentivized decisions in a multiple-price list to elicit risk preferences ([26];"Risk" $\in [0, 10]$ where 10 is very risk-loving). Using data from the immediately preceding round, we create variables to track a participant's earnings ("P.Earn" is an integer between -1 and 28) and whether a participant was in an accident ("P.Acc"= 1). We also create a dummy variable tracking whether a choice is made in an early or late round of the session ("Late"= 1 in rounds after 10). We define $\mathbf{X}$ as the vector of 2 participant-specific variables (Female and Risk) and 3 round-specific variables (P.Earn, P.Acc, and Late) for which we control. In addition, we define $\mathbf{Z}$ as a vector containing all possible interactions between the model's condition variables and $\mathbf{X}$.[6]

We use this set of independent variables to explain the dependent variable of driving choice (which is either Fast, Slow, or Auto). We address the following 2 questions.

Question (1) If the presence of a policy deters Fast drivers, then which non-Fast action is chosen by these deterred drivers?

Question (2) If specific policies deter Fast drivers, then which non-Fast action is chosen by these deterred drivers in each policy?

[5] All 6 comparisons are significantly different from their respective *Control* at the $p < 0.001$ level (Pearson's Chi-Squared). This level of significance is also observed when comparing AnyPolicy and *Control*.

[6] For example, when using the AnyPolicy variable, as in model (1), $\mathbf{Z}$ contains 5 interaction terms (AnyPolicy*Female, AnyPolicy*Risk, AnyPolicy*P.Earn, AnyPolicy*P.Acc, and AnyPolicy*Late). In other models, such as model (2), $\mathbf{Z}$ contains 15 interaction variables.

We employ a multinomial logistic regression which, for each model, conducts 2 independent binary logistic regressions in which the Fast driving choice is used as a reference for which the other Slow and Auto are regressed against. As shown below, model (1) uses the pooled "AnyPolicy" independent variable whereas model (2) separately identifies each policy condition.

Model (1)

$$\ln\left(\frac{p(\text{Slow})}{p(\text{Fast})}\right) = \text{AnyPolicy} \cdot \beta_{1,S} + \mathbf{X}\beta_{\mathbf{X},S} + \mathbf{Z}\beta_{\mathbf{Z},S} + \beta_{0,S}$$

$$\ln\left(\frac{p(\text{Auto})}{p(\text{Fast})}\right) = \text{AnyPolicy} \cdot \beta_{1,A} + \mathbf{X}\beta_{\mathbf{X},A} + \mathbf{Z}\beta_{\mathbf{Z},A} + \beta_{0,A}$$

Model (2)

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

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

Table 4 presents the maximum likelihood estimates for the relevant variables.[7] Columns (1) and (2) show the estimates of models (1) and (2) which address questions (1) and (2).

[7]We use the 'mlogit' function with the "vce" option in Stata. Since observations are independent across sessions (but not within sessions), errors are clustered at the session level. Participants-level fixed effects are not included because each participant experiences only one condition. Tables 9, 10, and 11 report on multinomial logit models using the same Stata options. We find similar results to that shown in Table 4 in a model using a binary logistic regression where the dependent variable is Fast (1) or either of the non-Fast options (0). Estimations of this logit model are in Appendix C. In addition, Table 17 in Appendix D presents the estimates of all variables in models (1) and (2).

Table 4: Driving choice (relative to Fast)

	Slow (1a)	Auto (1b)	Slow (2a)	Auto (2b)
AnyPolicy	0.471	0.205	-	-
	(0.60)	(0.40)		
AnyPolicy*Female	0.516	0.806**	-	-
	(1.64)	(2.92)		
Endogenous	-	-	0.721	0.407
			(0.79)	(0.70)
*Endogenous**Female	-	-	0.895*	0.754*
			(2.20)	(2.27)
Exogenous	-	-	0.499	0.642
			(0.55)	(1.02)
*Exogenous**Female	-	-	0.187	0.615
			(0.47)	(1.56)
Framing	-	-	0.244	-0.415
			(0.30)	(-0.63)
*Framing**Female	-	-	0.408	1.078*
			(1.22)	(2.43)
Female	0.184	-0.221	0.184	-0.221
	(0.73)	(-1.16)	(-0.73)	(-1.16)
Risk	0.168	-0.140	0.168	-0.140
	(0.99)	(-1.45)	(0.99)	(-1.45)
P.Earn	-0.080***	-0.104***	-0.080***	-0.104***
	(-7.02)	(-11.84)	(-7.02)	(-11.84)
P.Acc	-1.626***	-2.173***	-1.626***	-2.173***
	(-6.67)	(-7.45)	(-6.67)	(-7.45)
Late	-0.0689	0.301***	-0.0689	0.301***
	(-0.54)	(6.40)	(-0.54)	(6.40)
{*Condition*}* {Risk, P.Earn, P.Acc, Late}	✓	✓	✓	✓
N	6749	6749	6749	6749
Pseudo-R^2	0.109		0.113	

t statistics in parentheses

[a] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Columns (1a) and (1b) show that female participants are more likely to choose non-Fast driving choices in the presence of a policy. Relative to the Fast driving choice, female participants are more likely to choose Auto (p = 0.003) and are marginally more likely to choose Slow (p = 0.101) in the presence of a policy condition. Furthermore, columns (2a) and (2b) show that this responsiveness is mostly present in *Endogenous* and *Framing*. In both *Endogenous* and *Framing*, female participants are more likely to shift from Fast into Auto (p = 0.024 and 0.015, respectively). In addition, in *Endogenous*, female participants are likely to shift from Fast into Slow (p = 0.028). The magnitude of these effects are displayed in log odds. Compared to male participants in *Control*, female participants in *Endogenous* have a 0.895 increase in the log odds of choosing Slow (relative to Fast) and a 0.754 increase in the log odds of choosing Auto (relative to Fast). Models (1) and (2) show that male participants do not change their driving choices in the presence of any policy. Estimates of AnyPolicy in columns (1a) and (1b) as well as estimates of *Endogenous*, *Exogenous*, and *Framing* in columns (2a) and (2b) all have p-values greater than 0.308.

Result (3)

In the presence of any policy condition, female participants are more likely choose Auto (relative to Fast). This is particularly salient in *Endogenous* and *Framing*. In addition, female participants are more likely choose Slow (relative to Fast) in *Endogenous*.

Female participants in *Control* are not more likely to choose either non-Fast option than male participants. Risk preference in *Control* does not explain driving choices and all interaction variables with risk-preference are insignificant at levels higher than 0.10. However, P.Earn and P.Acc are highly significant in predicting driving choices in *Control*. This is unsurprising since our model and experiment assumes higher expected earnings and higher accident probabilities with Fast drivers.[8] Finally, participants prefer Auto (relative to Fast) in late rounds.

5.3 Beliefs about the driving choices of others

In each round, a participant submits a belief about the proportion of Fast, Slow, and Auto drivers that will be present in the group. The accuracy of this elicited belief determines the amount of earnings in that round. The average earnings from these belief elicitations

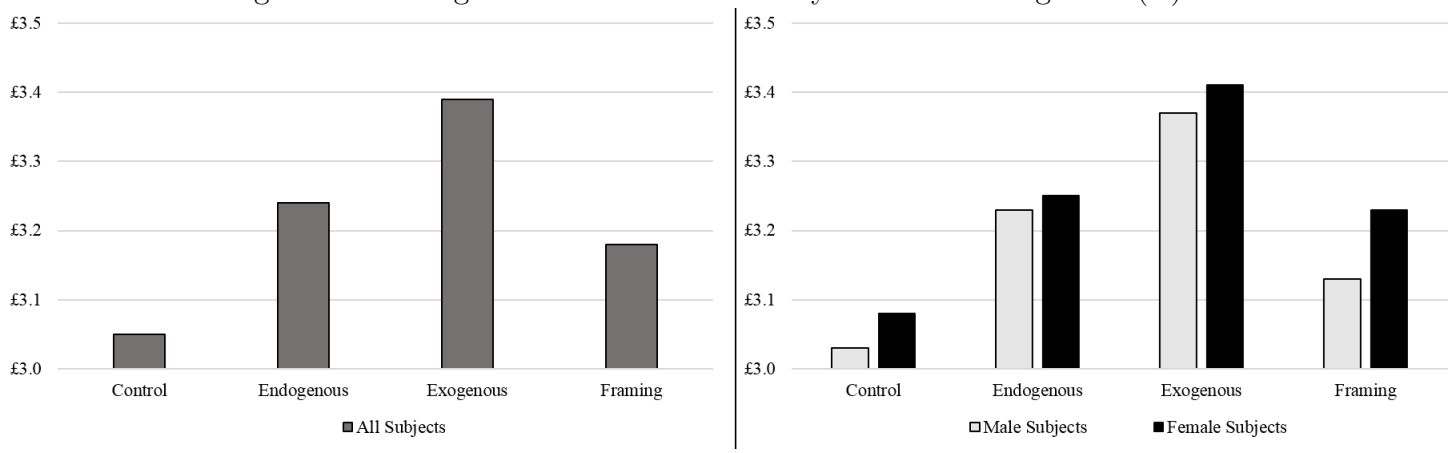
[8]There is no systematic difference in this narrative within any of the 3 policy conditions. Because of this, the interaction variables are omitted from Table 4. Table 17 presents the estimation of all variables in models (1) and (2).

separated by condition and gender are shown in Table 5 and Figure 4.

Table 5: Earnings from belief elicitation by condition and gender (£)

	Control	*Endogenous*	*Exogenous*	*Framing*
All participants	3.05	3.24	3.39	3.18
Male participants	3.03	3.23	3.37	3.13
Female participants	3.08	3.25	3.41	3.23
Male - Female	-0.05	-0.02	-0.04	-0.10
Diff (p-value)	.362	.620	.449	.066

Figure 4: Earnings from belief elicitation by condition and gender (£)



When analyzing all participants (top row of Table 5 and left panel of Figure 4), each policy condition produces significantly higher payoffs relative to *Control* ($p< 0.001$ using a two-sample t-test for all 3 pairwise comparisons). This means that participants are more accurate at predicting the driving choices of others in the presence of any policy condition. Furthermore, earnings from belief elicitations are not different across gender in *Control*, *Endogenous*, or *Exogenous*. Female participants are marginally more accurate in *Framing* ($p= 0.066$). This suggests that inaccurate beliefs of female participants cannot explain the observed differences in AS, driving choice payoffs, or driving choices (Results 1, 2, and 3). In fact, female participants are marginally more accurate than male participants at predicting the choices of others, which makes their driving choices even more puzzling.

As with driving choices, we use the previously described set of independent variables (condition type, **X**, and **Z**) to explore the relationship between beliefs and policies. We address the following 2 questions.

Question (3) Does the presence of a policy change beliefs about Fast/Slow/Auto drivers in the population?

Question (4) Which specific policy changes the beliefs about Fast/Slow/Auto drivers in the population?

For models (3) and (4), we employ 3 separate linear regressions to address each question. Model (3) uses the pooled "AnyPolicy" independent variable whereas model (4) separately identifies each policy condition. Table 6 presents the maximum likelihood estimates for the relevant variables.[9] As with the analysis on driving choices, the column number aligns with the model number and question number.

[9]We use the 'reg' function with the "vce" option in Stata. Table 18 in Appendix D presents the estimates of all variables in models (3) and (4).

Table 6: Beliefs about the driving choices in the population

	Fast Belief (3a)	Slow Belief (3b)	Auto Belief (3c)	Fast Belief (4a)	Slow Belief (4b)	Auto Belief (4c)
AnyPolicy	-1.124	-1.127	2.251	-	-	-
	(-0.29)	(-0.32)	(0.88)			
AnyPolicy*Female	-2.851	-1.768	4.618*	-	-	-
	(-1.51)	(-1.19)	(2.19)			
Endogenous	-	-	-	-2.528	1.219	1.309
				(-0.59)	(0.32)	(0.49)
*Endogenous**Female	-	-	-	-4.995*	-0.827	5.821*
				(-2.49)	(-0.56)	(2.59)
Exogenous	-	-	-	3.286	-5.934[a]	2.648
				(0.72)	(-1.75)	(0.63)
*Exogenous**Female	-	-	-	-3.963	-0.715	4.678
				(-1.62)	(-0.30)	(1.37)
Framing	-	-	-	-1.991	0.128	1.863
				(-0.41)	(0.02)	(0.41)
*Framing**Female	-	-	-	0.309	-3.756*	3.447[a]
				(0.18)	(-2.42)	(1.79)
Female	1.395	2.848*	-4.243*	1.395	2.848*	-4.243*
	(0.86)	(2.73)	(-2.56)	(0.85)	(2.73)	(-2.56)
Risk	-0.0546	-0.102	0.157	-0.0548	-0.102	0.157
	(-0.09)	(-0.23)	(0.32)	(-0.09)	(-0.23)	(0.32)
P.Earn	0.258**	-0.124*	-0.134	0.258**	-0.124*	-0.134
	(3.28)	(-2.58)	(-1.30)	(3.29)	(-2.58)	(-1.30)
P.Acc	4.767*	-1.488	-3.279	4.773*	-1.495	-3.278
	(2.71)	(-1.95)	(-1.62)	(2.71)	(-1.96)	(-1.61)
Late	0.557	-5.316***	4.759**	0.557	-5.316***	4.759**
	(0.36)	(-7.60)	(3.17)	(0.36)	(-7.60)	(3.17)
{*Condition*}* {Risk, P.Earn, P.Acc, Late}	✓	✓	✓	✓	✓	✓
N	6749	6749	6749	6749	6749	6749
R^2	0.049	0.059	0.015	0.074	0.078	0.056

t statistics in parentheses

[a] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Column (3c) shows that female participants facing a policy condition believe that the population will consist of more Auto drivers (p = 0.036). Furthermore, column (4c) shows that this effect is mostly present in *Endogenous* (p = 0.014).[10] Male participants do not change their beliefs about the proportion of Auto drivers in the presence of any policy. Estimates of AnyPolicy in column (3c) and estimates of *Endogenous*, *Exogenous*, and *Framing* in column (4c) all have p-values greater than 0.383.

Given that female participants believe Auto will be chosen more often in the presence of a policy, what driving choice do they believe will be chosen less often? Interestingly, this depends on the specific policy. Columns (4a) and (4b) show that female participants believe *Endogenous* reduces Fast drivers whereas *Framing* reduces Slow drivers (p = 0.018 and 0.022, respectively). While male participants do show a marginally significant decrease in their belief about Slow drivers in *Exogenous* (p= 0.090), a systematic change does not exist in the presence of any policy. Estimates of AnyPolicy in columns (3a) and (3b) as well as estimates of *Endogenous*, *Exogenous*, and *Framing* in columns (4a) and (4b) all have p-values greater than 0.478. This analysis is combined into our next result.

Result (4)

In the presence of any policy, particularly *Endogenous*, female participants believe others are more likely to choose Auto. In addition, female participants believe others are less likely to choose Fast in *Endogenous* whereas, in *Framing*, female participants believe others are less likely to choose Slow.

Female participants in *Control*, when compared to their male counterparts, believe that others are less likely to select Auto and more likely to select Slow. Risk preference in *Control* does not explain beliefs and interactions with risk-preference are not consistently significant. If a participant experiences a higher payoff or an accident in the previous round, they believe that others are more likely to choose Fast. Finally, participants in *Control* believe that others are less likely to choose Slow and more likely to choose Auto in late rounds.[11]

5.4 Social norms as a mechanism

After the the driving rounds, we elicit each participant's first-order and second-order normative beliefs about the social norm within the population. Participants are asked the following

[10] The effect is marginally present in *Framing* (p = 0.082).

[11] Many of the the interaction variables are omitted from Table 6. Table 18 presents the estimation of all variables in models (3) and (4).

question.

Question (1): In general, what do you think another participant in the experiment should do in this situation?

Their answer to Question 1 (either Fast, Slow or Auto) represents that participant's first-order normative belief about the social norm in the population. After submitting their answer, participants are asked this follow-up question.

Question (2): In general, what do you think others believe another participant in the experiment should have done? (This is the same as asking 'what driving method do you think most people in the room chose to answer Question (1)')

Their answer to Question 2 (either Fast, Slow or Auto) represents that participant's second-order normative belief about the social norm in the population.[12] The percentage of participants who state Fast or Auto as their first-order beliefs are shown in Table 7 and Figure 5.[13]

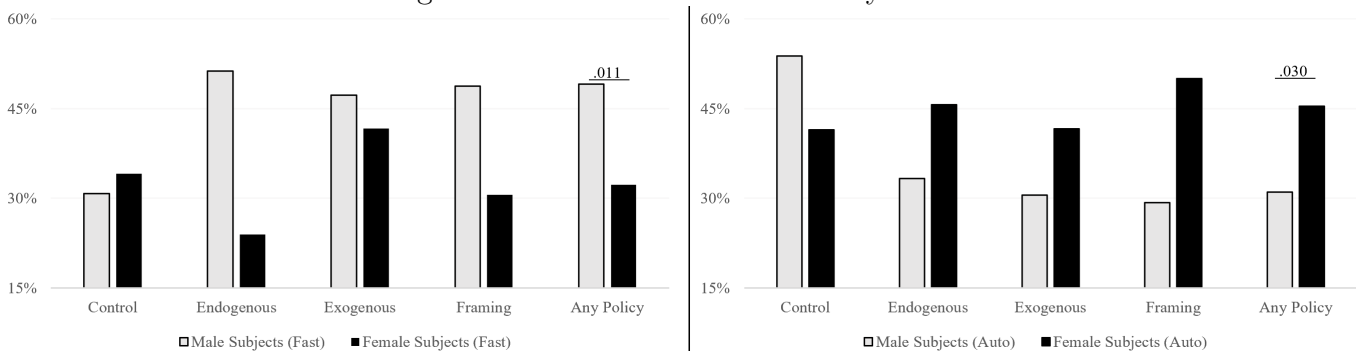
Table 7: First-order social norm by condition

	% Fast				*% Auto*			
Condition	Male	Female	Male - Female	Diff (p-value)	Male	Female	Male - Female	Diff (p-value)
Control	30.8	34.1	-3.3	0.933	53.8	41.5	12.3	0.376
Framing	48.8	30.6	18.2	0.163	29.3	50.0	-20.7	0.104
Exogenous	47.2	41.7	5.5	0.775	30.6	41.7	-11.1	0.415
Endogenous	51.3	23.9	27.4	0.017	33.3	45.7	-12.4	0.351
AnyPolicy	49.1	32.3	16.8	0.011	31.0	45.4	-14.4	0.030

[12]Question (1) is not incentivized. After submitting the answer to Question (1), participants are told that if their answer to Question (2) was the same as the most-chosen answer to Question (1) within their population, they would earn £2.

[13]Slow is omitted because it is the least chosen option and because the percentage of participants who choose Slow are not statistically different across gender.

Figure 5: First-order social norm by condition



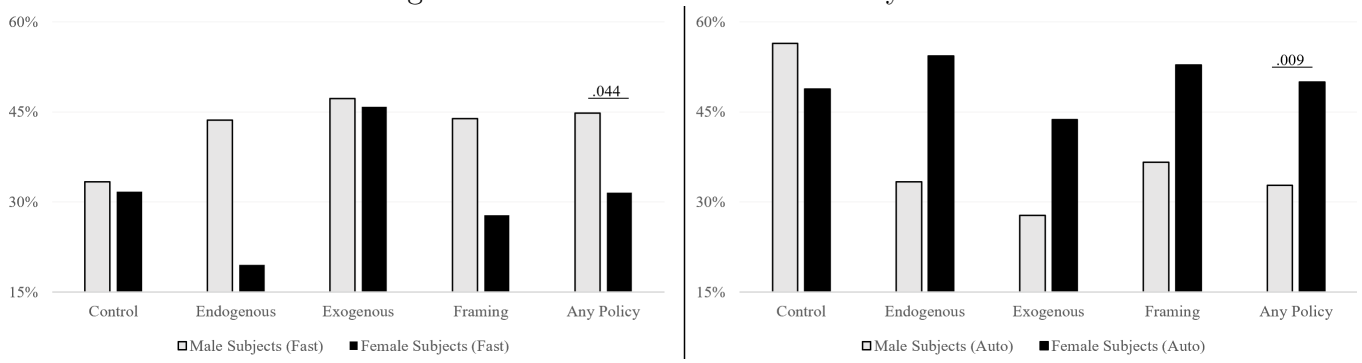
In *Control*, the percentage of participants who state Fast as their first-order social norm is not different across gender (p= 0.933). The same is true for stating Auto as the first-order social norm (p= 0.376). In the presence of any policy, compared to their male counterparts, female participants are less likely to state Fast and more likely to state Auto as the first-order social norm ($p = 0.011$ and p= 0.030 using a two-sample chi-squared test comparing Fast with non-Fast and Auto with non-Auto). This effect is largely driven by the behavior in *Endogenous* and *Framing*. More specifically, compared to male participants, Fast is chosen significantly less often by female participants in *Endogenous* (p=0.017) and Auto is chosen marginally more often by female participants in *Framing* (p=0.104).

The percentage of participants who state Fast or Auto as their second-order beliefs are shown in Table 8 and Figure 6.

Table 8: Second-order social norm by condition

	% Fast				*% Auto*			
Condition	Male	Female	Male - Female	Diff (p-value)	Male	Female	Male - Female	Diff (p-value)
Control	33.3	31.7	1.6	1.000	56.4	48.8	7.6	0.646
Framing	43.9	27.8	16.1	0.219	36.6	52.8	-16.2	0.231
Exogenous	47.2	45.8	1.4	1.000	27.8	43.8	-16.0	0.203
Endogenous	43.6	19.6	24.0	0.031	33.3	54.3	-21.0	0.055
AnyPolicy	44.8	31.5	13.3	0.044	32.8	50.0	-17.2	0.009

Figure 6: Second-order social norm by condition



In *Control*, the percentage of participants who state Fast as their second-order social norm is not different across gender (p= 1.000). The same is true for stating Auto as the second-order social norm (p= 0.646). In the presence of any policy, compared to their male counterparts, female participants are less likely to state Fast and more likely to state Auto as the first-order social norm (p= 0.044 and p= 0.009 using a two-sample chi-squared test comparing Fast with non-Fast and Auto with non-Auto). This effect is largely driven by behavior in *Endogenous*. More specifically, compared to male participants, Fast is chosen significantly less often by female participants (p= 0.031) and Auto is chosen significantly more often by female participants in *Endogenous* (p=0.055).

Both first-order and second-order social norms consistently differ across gender in the presence of a policy. As with driving choices and belief elicitations, we aim explore the relationship between these social norms and policies. We address the following 2 questions.

Question (5) If the presence of a policy deters the Fast-driving norm, then which non-Fast norm prevails?

Question (6) If specific policies deter the Fast-driving norm, then which non-Fast norm prevails in each policy?

For models (5) and (6), we employ multinomial logistic regressions similar to models (1) and (2). The main difference from those models is that the dependent variable is the norm choice rather than driving choice. In addition, since we only have one data point for each participant, round-specific variables are not included. Tables 9 and 10 present the maximum likelihood estimates of models (5) and (6) for the first-order and second-order beliefs, respectively.

Table 9: First-order social norm (relative to Fast)

	Slow (5a.1)	Auto (5b.1)	Slow (6a.1)	Auto (6b.1)
AnyPolicy	0.246	-0.061	-	-
	(0.14)	(-0.08)		
AnyPolicy*Female	0.175	1.073[a]	-	-
	(0.24)	(1.92)		
Endogenous	-	-	0.333	1.434
			(0.15)	(1.36)
*Endogenous**Female	-	-	1.078	1.501*
			(1.38)	(2.30)
Exogenous	-	-	0.300	-0.162
			(0.15)	(-0.14)
*Exogenous**Female	-	-	-0.528	0.565
			(-0.61)	(0.80)
Framing	-	-	0.280	-0.795
			(0.15)	(-0.86)
*Framing**Female	-	-	-0.004	1.346[a]
			(0.01)	(1.88)
Female	0.352	-0.341	0.352	-0.341
	(0.52)	(-0.73)	(0.52)	(-0.73)
Risk	0.065	-0.126	0.065	-0.126
	(0.20)	(-1.23)	(0.20)	(-1.23)
AnyPolicy*Risk	-0.107	-0.220	-	-
	(-0.31)	(-1.35)		
*Endogenous**Risk	-	-	-0.195	-0.641**
			(-0.44)	(-3.23)
*Exogenous**Risk	-	-	-0.083	-0.149
			(-0.23)	(0.59)
*Framing**Risk	-	-	-0.093	-0.063
			(-0.25)	(-0.37)
Constant	-0.961	1.039[a]	-0.961	1.039[a]
	(-0.56)	(1.76)	(-0.56)	(1.76)
N	326	326	326	326
Pseudo-R^2	0.041		0.057	

t statistics in parentheses

[a] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 10: Second-order social norm (relative to Fast)

	Slow (5a.2)	Auto (5b.2)	Slow (6a.2)	Auto (6b.2)
AnyPolicy	0.408	-0.962	-	-
	(1.37)	(-1.22)		
AnyPolicy*Female	-0.503	0.794	-	-
	(-0.86)	(1.35)		
Endogenous	-	-	0.152	0.979
			(0.08)	(1.02)
*Endogenous**Female	-	-	0.254	1.473*
			(0.37)	(2.20)
Exogenous	-	-	0.308	-0.930
			(0.12)	(-0.87)
*Exogenous**Female	-	-	-1.436	0.401
			(-1.46)	(0.51)
Framing	-	-	0.475	-2.235**
			(0.35)	(-2.79)
*Framing**Female	-	-	-0.218	0.917
			(-0.31)	(1.35)
Female	0.698	-0.058	0.698	-0.058
	(1.63)	(-0.11)	(1.63)	(-0.11)
Risk	0.129	-0.210[a]	0.129	-0.210[a]
	(0.43)	(-1.74)	(0.43)	(-1.74)
AnyPolicy*Risk	0.013	0.046	-	-
	(0.04)	(0.30)		
*Endogenous**Risk	-	-	0.081	-0.489*
			(0.18)	(-2.27)
*Exogenous**Risk	-	-	0.027	0.012
			(0.06)	(0.07)
*Framing**Risk	-	-	-0.015	0.387*
			(-0.05)	(2.59)
Constant	-1.744	1.330[a]	-1.744	1.330[a]
	(-1.49)	(1.96)	(-1.49)	(1.96)
N	326	326	326	326
Pseudo-R^2	0.035		0.071	

t statistics in parentheses

[a] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

For the first-order beliefs about social norms, column (5b.1) shows that female participants in the presence of a policy are more likely to state Auto as the social norm (relative to Fast; p = 0.055). Furthermore, column (6b.1) shows that this responsiveness is mostly present in *Endogenous* and *Framing* (p = 0.021 and 0.060, respectively). In summary, the effect of policies on first-order social norms mimics the effect of policies on driving choices (Result 3). Therefore our Result 5, here, aligns with Result 3.

Result (5)

In the presence of any policy condition, female participants are more likely state Auto as their first-order social norm instead of Fast. This is particularly salient in *Endogenous* and *Framing*.

For the second-order beliefs about social norms, we don't observe a statistically significant effect of the presence of a policy. However, column (6b.2) shows that female participants in *Endogenous* are more likely to state Auto (as opposed to Fast) as the social norm whereas male participants are less likely to state Auto (as opposed to Fast) in *Framing* (p = 0.028 and 0.005, respectively).

Result (6)

Female participants are more likely state Auto as their second-order social norm instead of Fast in *Endogenous*. Male participants are less likely state Auto as their second-order social norm instead of Fast in *Framing*.

Results 5 and 6 suggest that female participants create stronger social norms around non-Fast driving choices (especially in *Endogenous* and *Framing*). If this is true, then one would further expect that female participants would be more responsive to realizing a violation of their (non-Fast) norm. Do female participants, indeed, respond differently to the realization of a speeding fine? We test this by analyzing a participant's driving choice before and after the realization of a driving fine. Since this analysis requires the realization of a driving fine, we use the 96 participants from *Endogenous* and *Exogenous* who realize a driving fine at some point during their session (47 participants from *Exogenous* and 49 from *Endogenous*).[14] We address the following 2 questions.

Question (7) Does the realization of a driving fine in round t coincide with more non-Fast driving choices in rounds $t+1$, $t+2$, ... T?

[14]There are 169 total participants in both conditions and 73 participants never receive a fine.

Question (8) If realizing a driving fine in round t deters Fast driving choices in rounds $t+1$, $t+2$, ... T, then which non-Fast driving choice is chosen?

Model (7) employs a binary logistic regression where the dependent variable equals either 0 (Fast) or 1 (non-Fast). Model (8) employs a multinomial logistic regression similar to model (1) where the main difference is the inclusion of a dummy variable ("AfterFine") which equals 1 for all rounds after a participant has realized a driving fine. Models (7) and (8) use $\mathbf{X}$ and $\mathbf{Z}$ as previously described and Table 11 presents maximum likelihood estimates for the relevant variables.[15]

Table 11: Driving choice (relative to Fast) - Response to first experienced fine

	Logit	*Multinomial Logit*	
	Non-Fast	Slow	Auto
	(7)	(8a)	(8b)
AfterFine	0.129	-0.728	0.670
	(0.14)	(-0.69)	(0.65)
AfterFine*Female	0.698^{a}	0.649^{a}	0.878^{a}
	(1.94)	(1.66)	(1.73)
Female	-0.016	0.282	-0.391
	(-0.06)	(0.80)	(-1.16)
Risk	-0.087	-0.061	-0.120
	(-0.84)	(-0.55)	(-0.94)
P.Earn	-0.030***	-0.019*	-0.045***
	(-4.13)	(-2.35)	(-3.64)
P.Acc	-0.652***	-0.411^{a}	-0.945***
	(-3.55)	(-1.89)	(-4.48)
Late	0.085	-0.181	0.385**
	(0.41)	(-0.58)	(2.70)
AfterFine* {Risk, P.Earn, P.Acc, Late}	✓	✓	✓
N	2042	2042	2042
Pseudo-R^2	0.073	0.065	

t statistics in parentheses

^{a} $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

[15]Table 19 in Appendix D presents the estimates of all variables in models (7) and (8).

Column (7) shows that female participants are more likely to choose a non-Fast driving choice after experiencing a fine ($p = 0.052$). Furthermore, columns (8a) and (8b) show that, relative to the Fast driving choice, female participants are more likely to choose Auto ($p = 0.083$) and Slow ($p = 0.097$) after realizing a driving fine. Models (7) and (8) shows that male participants do not change their driving choices after realizing a driving fine. Estimates of AfterPen in columns (7), (8a), and (8b) have respective p-values of 0.890, 0.491, and 0.517.

Result (7)

Female participants are more likely to choose non-Fast driving choices after realizing a driving fine.

6 Conclusion

In this paper, we have studied which policies can reduce driving speeds when different driving styles coexist — 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. A first conclusion of our research is that neither of these three scenarios leads to a reduction in average speed, which is the main factor influencing the risk of vehicle accidents involving injuries. Thus our first contribution is to point out the difficulty of designing mechanisms that lead to safer driving, or more generally reducing socially harmful behavior. A second, and possibly more important, insight is that policies have heterogeneous effects in the population. In particular the average effect on genders is markedly different.

Indeed, the lack of a global average effect on speed and other relevant magnitudes arises from changes of opposite sign in male and female participants' behavior in response to the proposed policies. In particular, average speed remains the same while female participants reduce their average speed, while male participants increase it. In our experiment, this goes along with a distributional effect regarding the participants' earnings: the overall average earnings do not change in the different policy conditions, but men increase their earnings at the expense of female participants when policies are implemented. This means that the differences in gender reaction to policies is not without consequences, and female participants are harmed by their more prosocial choices. The choices of female participants are actually the ones we expected a priori based on theoretical considerations: in the presence of policy conditions, female participants choose Auto instead of Fast. While this is particularly true

in *Endogenous* and *Framing*, we have observed that in *Endogenous* female participants also choose Slow instead of Fast, suggesting that *Endogenous* is the most effective policy among the ones tested.

We also establish the likely mechanisms for our results. To this we use data from beliefs about others' behavior and about others' "proper" behavior. We also use the reactions of different participants to fines, and their willingness to contribute to fines. All those data point to the fact that in the presence of any policy, the social norm of male participants is more likely to be Fast than the norm of female participants. As a result, female participants reduce their use of Fast driving styles, which makes it more profitable for males to use Fast, thereby annulling the effect of the policy.

In summary, there are two main take away message from our experimental results. One is that policies aimed at promoting prosocial behavior have no aggregate effect. The other is that effects differ across genders. More specifically, these policies are effective at changing female participants' driving behavior and female participants' beliefs about the driving behavior of others, but male participants have an opposing reaction. This is very important for policy. Not only might policies be ineffective, but the behavioral reactions in the population can increase inequity. We already know that behavioral reactions can mitigate the effect of policies (as in the example of seatbelt legislation [1]), but this complete cancellation and worsening of inequality is more worrying. It may also lead to reduced contributions in a dynamic environment. We have not explored this latter impact, as the socio-demographic characteristics of the participants are not communicated to others in the experiment, but it is an important consideration for future research.

We have also shown that policy effects appear to be mediated by social norms which are more prevalent among female drivers. Therefore, our results suggest that a proper policy to prepare for mixed-agency scenarios is through behavior change interventions that appeal directly to peoples expectations about what others will do ([11]), and that can be preferentially addressed to men. In connection with this, further research from other disciplines such as neuroscience and psychology may help provide and understanding for the underlying mechanisms which may explain our differences found between the genders. Overall, our findings suggest that future research which analyzes driving behavior, or more generally any prosocial behavior, should remain vigilant about the possibility of gender differences in this context.

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 10.4 (2018): 159-89.

[3] Andreoni, James, and Lise Vesterlund. "Which is the fair sex? Gender differences in altruism." The Quarterly Journal of Economics 116.1 (2001): 293-312.

[4] Ben-Ner, Avner, Fanmin Kong, and Louis Putterman. "Share and share alike? Gender-pairing, personality, and cognitive ability as determinants of giving." Journal of Economic Psychology 25.5 (2004): 581-589.

[5] Brañas-Garza, Pablo, Valerio Capraro and Ericka Rascón-Ramírez. "Gender differences in altruism on Mechanical Turk: Expectations and actual behaviour". Economics Letters 170 (2018): 19–23.

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

[7] Bigoni, Maria, Gabriele Camera, and Marco Casari. "Partners or Strangers? Cooperation, monetary trade, and the choice of scale of interaction." American Economic Journal: Microeconomics 11.2 (2019): 195-227.

[8] 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.

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

[10] Chen, Wanting, Shuyue Zhang, Ofir Turel, Youqing Peng, Hong Chen, Qinghua He "Sex-based differences in right dorsolateral prefrontal cortex roles in fairness norm compliance." Behavioural brain research 361 (2019): 104-112.

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

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

[13] Cooper, Russell, Douglas V. DeJong, Robert Forsythe, and Thomas W. Ross. "Cooperation without reputation: Experimental evidence from prisoner's dilemma games." Games and Economic Behavior 12, no. 2 (1996): 187-218.

[14] Croson, Rachel, and Uri Gneezy. "Gender differences in preferences." Journal of Economic literature 47.2 (2009): 448-74.

[15] 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.

[16] Dreber, Anna and Johanesson, Magnus. "Gender differences in deception". Economics Letters 99 (2008): 197–199.

[17] Eckel, Catherine C., and Philip J. Grossman. "Differences in the economic decisions of men and women: Experimental evidence." Handbook of experimental economics results 1 (2008): 509-519.

[18] Eckel, Catherine C., and Philip J. Grossman. "The relative price of fairness: Gender differences in a punishment game." Journal of Economic Behavior & Organization 30.2 (1996): 143-158.

[19] 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.

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

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

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

[23] Grosch, Kerstin and Holger A. Rau. "Gender differences in compliance: The role of social value orientation", GlobalFood Discussion Papers, No. 88, Georg-August- Universität Göttingen, Research Training Group (RTG) 1666 - GlobalFood, Göttingen (2016).

[24] 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.

[25] Herrmann, Benedikt, Christian Thöni, and Simon Gächter. "Antisocial punishment across societies." Science 319, no. 5868 (2008): 1362-1367.

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

[27] 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.

[28] 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.

[29] Lan, Tian and Ying-yi Hong. "Norm, gender, and bribe-giving: Insights from a behavioral game." PLoS ONE 12 (2017): e0189995.

[30] 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.

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

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

[33] Özkan, Türker, and Timo Lajunen. "What causes the differences in driving between young men and women? The effects of gender roles and sex on young drivers driving behaviour and self-assessment of skills." Transportation research part F: Traffic psychology and behaviour 9.4 (2006): 269-277.

[34] 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.

[35] "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.

[36] Solow, John L., and Nicole Kirkwood. "Group identity and gender in public goods experiments." Journal of Economic Behavior & Organization 48.4 (2002): 403-412.

[37] Ulleberg, Pål. "Social influence from the back-seat: factors related to adolescent passengers willingness to address unsafe drivers." Transportation Research Part F: Traffic Psychology and Behaviour 7.1 (2004): 17-30.

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

[39] Weatherburn, Don, and Steve Moffatt. "The specific deterrent effect of higher fines on drink-driving offenders." The British Journal of Criminology 51.5 (2011): 789-803.

[40] "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.

[41] "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.

[42] "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.

[43] "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 $\frac{(1-0.3AS)^2}{(1-0.35AS)} < 1$ 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

$$\frac{((1-p)\,U_i\,(S_F)+pU_i\,((S_F-P)))}{U_i\,(S_A)}=\frac{\left(1-a_AAS_i^P\right)}{(1-a_FAS_i^P)}. \tag{5}$$

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

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

This means that the following must be true:

$$\frac{(1-a_AAS_i)}{(1-a_FAS_i)}>\frac{\left(1-a_AAS_i^P\right)}{(1-a_FAS_i^P)}$$

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

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

Note that $AS=x_FS_F+x_SS_S+x_AS_A$. 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 - Participants statistics and Balance Check

Table 12 shows the session characteristics divided by condition.

Table 13 compares the participant characteristics divided by condition. In addition to gender and risk preference, participants are asked their age, whether they had ever owned a driver's license ("License"), and whether they learned during the experiment ("Learning"). P-values are shown for differences (from *Control*) using either a chi-squared test across measures using proportions (Female, License, and Learning) or a Mann-Whitney test across measures with ordinal data (Risk and Age).

Tables 14 and 15 compare the participant characteristics divided by condition and by gender. P-values are shown for differences (from *Control*) using either a chi-squared test

Table 12: Session statistics by policy condition

Condition	# of participants	Avg. time of day	Avg. # of rounds	# of choices
Control	80	12:48	21.6	1,735
Framing	77	12:15	18.9	1,457
Exogenous	84	12:33	20.4	1,703
Endogenous	85	12:48	21.8	1,854
Total	326	12:36	20.7	6,749

Table 13: Participants statistics by policy condition

Condition	Female (prop.)	Avg. Risk	Avg. Age	License (prop.)	Learning (prop.)
Control	.51	3.94	24.2	.65	.60
Framing	.47	4.03	24.4	.71	.56
	($p = .69$)	($p = .59$)	($p = .48$)	($p = .49$)	($p = .71$)
Exogenous	.57	4.40	24.4	.64	.62
	($p = .55$)	($p = .06$)	($p = .61$)	($p = 1.0$)	($p = .93$)
Endogenous	.54	3.92	23.8	.72	.67
	($p = .83$)	($p = .88$)	($p = .90$)	($p = .44$)	($p = .44$)
AnyPolicy	.53	4.12	24.2	.69	.62
	($p = .91$)	($p = .36$)	($p = .66$)	($p = .58$)	($p = .88$)

(License and Learning) or a Mann-Whitney test (Risk and Age), depending on the data.

Out of the 32 tests performed in Tables 14 and 15, only one is significantly different. Thus, the conditions are well-balanced across gender.

Appendix C - Additional analysis

Here we employ a binary logistic regression to address the following 2 questions.

Question (C.1) Does the presence of a policy deter Fast drivers?

Question (C.2) Which specific policies deter Fast drivers?

The dependent variable is either 0 (Fast) or 1 (either Slow or Auto). As shown below, model (C.1) uses the pooled "AnyPolicy" variable whereas model (C.2) separately identifies each policy condition.

Table 14: Participants statistics by policy condition - Female participants

Condition	Avg. Risk	Avg. Age	License (prop.)	Learning (prop.)
Control	4.05	24.1	.61	.68
Framing	3.97	24.0	.72	.53
	($p = .86$)	($p = .60$)	($p = .42$)	($p = .25$)
Exogenous	4.04	24.7	.71	.58
	($p = .93$)	($p = .55$)	($p = .45$)	($p = .45$)
Endogenous	3.87	24.1	.70	.72
	($p = .46$)	($p = .50$)	($p = .54$)	($p = .91$)
AnyPolicy	3.96	24.3	.71	.62
	($p = .84$)	($p = .46$)	($p = .33$)	($p = .55$)

Table 15: Participants statistics by policy condition - Male participants

Condition	Avg. Risk	Avg. Age	License (prop.)	Learning (prop.)
Control	3.82	24.2	.69	.51
Framing	4.07	24.7	.71	.59
	($p = .49$)	($p = .63$)	($p = 1.0$)	($p = .67$)
Exogenous	4.89	24.1	.56	.67
	($p = .01$)	($p = .86$)	($p = .32$)	($p = .26$)
Endogenous	3.97	23.5	.74	.62
	($p = .66$)	($p = .42$)	($p = .80$)	($p = .49$)
AnyPolicy	4.29	24.1	.67	.62
	($p = .15$)	($p = .96$)	($p = .97$)	($p = .32$)

Model (C.1)

$$\ln\left(\frac{p(\text{Slow or Auto})}{p(\text{Fast})}\right) = \text{AnyPolicy} \cdot \beta_1 + \mathbf{X}\beta_{\mathbf{X}} + \mathbf{Z}\beta_{\mathbf{Z}} + \beta_0$$

Model (C.2)

$$\ln\left(\frac{p(\text{Slow or Auto})}{p(\text{Fast})}\right) = Endogenous \cdot \beta_1 + Exogenous \cdot \beta_2 + Framing \cdot \beta_3 + \mathbf{X}\beta_{\mathbf{X}} + \mathbf{Z}\beta_{\mathbf{Z}} + \beta_0$$

Table 16 presents maximum likelihood estimates where the column number aligns with the question and model number assigned above. These estimations are put into a table to be compared with models 1 and 2 in Table 16 [16]

[16] We use the 'logit' function with the "vce" option in Stata. Since observations are independent across sessions (but not within sessions), errors are clustered at the session level. Participants-level fixed effects are not included because each participant experiences only one condition.

Table 16: Driving choice (relative to Fast)

	Logit		*Multinomial Logit*			
	Slow or Auto		Slow	Auto	Slow	Auto
	(C.1)	(C.2)	(1a)	(1b)	(2a)	(2b)
AnyPolicy	0.179	-	0.471	0.205	-	-
	(0.36)		(0.60)	(0.40)		
AnyPolicy*Female	0.710***	-	0.516	0.806**	-	-
	(3.17)		(1.64)	(2.92)		
Endogenous	-	0.409	-	-	0.721	0.407
		(0.75)			(0.79)	(0.70)
*Endogenous**Female	-	0.856***	-	-	0.895*	0.754*
		(3.11)			(2.20)	(2.27)
Exogenous	-	-0.451	-	-	0.499	0.642
		(0.78)			(0.55)	(1.02)
*Exogenous**Female	-	0.442	-	-	0.187	0.615
		(1.42)			(0.47)	(1.56)
Framing	-	-0.244	-	-	0.244	-0.415
		(-0.43)			(0.30)	(-0.63)
*Framing**Female	-	0.825**	-	-	0.408	1.078*
		(2.50)			(1.22)	(2.43)
Female	-0.075	-0.075	0.184	-0.221	0.184	-0.221
	(-0.46)	(-0.46)	(-0.73)	(-1.16)	(-0.73)	(-1.16)
Risk	-0.025	-0.025	0.168	-0.140	0.168	-0.140
	(-0.22)	(-0.22)	(0.99)	(-1.45)	(0.99)	(-1.45)
P.Earn	-0.095***	-0.095***	-0.080***	-0.104***	-0.080***	-0.104***
	(-10.25)	(-10.25)	(-7.02)	(-11.84)	(-7.02)	(-11.84)
P.Acc	-1.970***	-1.970***	-1.626***	-2.173***	-1.626***	-2.173***
	(-7.84)	(-7.84)	(-6.67)	(-7.45)	(-6.67)	(-7.45)
Late	0.151**	0.151**	-0.0689	0.301***	-0.0689	0.301***
	(2.23)	(2.23)	(-0.54)	(6.40)	(-0.54)	(6.40)
{*Condition*}* {Risk, P.Earn, P.Acc, Late}	✓	✓	✓	✓	✓	✓
N	6749	6749	6749	6749	6749	6749
Pseudo-R^2	0.131	0.133	0.109		0.113	

t statistics in parentheses

[a] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Column (C.1) shows that female participants are more likely to choose non-Fast driving choices in the presence of a policy (p-value = .002). The magnitude of this effect is displayed in log odds. Compared to males in *Control*, female participants in any policy condition have a 0.710 increase in the log odds of choosing a non-Fast action. Furthermore, column (C.2) shows that policy-responsiveness is only present in *Endogenous* and *Framing* (p-value = .002 and .013, respectively). Male participants do not change their behavior in the presence of any policy. Estimates of AnyPolicy in column (C.1) and of *Endogenous*, *Exogenous*, and *Framing* in column (C.2) all have p-values greater than .433. This further supports the fact that, in the presence of any policy condition, female participants choose Fast less often which is particularly salient in *Endogenous* and *Framing*.

Appendix D - Full regression results

Below is the complete Table 4 including all variables. The table above the horizontal line corresponds to Table 4.

Table 17: Driving choice (relative to Fast)

	Slow (1a)	Auto (1b)	Slow (2a)	Auto (2b)
AnyPolicy	0.471	0.205	-	-
	(0.60)	(0.40)		
AnyPolicy*Female	0.516	0.806**	-	-
	(1.64)	(2.92)		
Endogenous	-	-	0.721	0.407
			(0.79)	(0.70)
*Endogenous**Female	-	-	0.895*	0.754*
			(2.20)	(2.27)
Exogenous	-	-	0.499	0.642
			(0.55)	(1.02)
*Exogenous**Female	-	-	0.187	0.615
			(0.47)	(1.56)
Framing	-	-	0.244	-0.415
			(0.30)	(-0.63)
*Framing**Female	-	-	0.408	1.078*
			(1.22)	(2.43)
Female	0.184	-0.221	0.184	-0.221
	(0.73)	(-1.16)	(-0.73)	(-1.16)
Risk	0.168	-0.140	0.168	-0.140
	(0.99)	(-1.45)	(0.99)	(-1.45)
P.Earn	-0.080***	-0.104***	-0.080***	-0.104***
	(-7.02)	(-11.84)	(-7.02)	(-11.84)
P.Acc	-1.626***	-2.173***	-1.626***	-2.173***
	(-6.67)	(-7.45)	(-6.67)	(-7.45)
Late	-0.0689	0.301***	-0.0689	0.301***
	(-0.54)	(6.40)	(-0.54)	(6.40)
AnyPolicy*Risk	-0.216	-0.130	-	-
	(-1.22)	(-1.16)		
AnyPolicy*P.Earn	0.022[a]	-0.007	-	-
	(1.68)	(-0.50)		
AnyPolicy*P.Acc	0.375	-0.203	-	-
	(1.29)	(-0.55)		
AnyPolicy*Late	-0.100	-0.076	-	-
	(-0.67)	(-1.00)		
*Endogenous**Risk	-	-	-0.298	-0.145
			(-1.57)	(-1.38)
*Endogenous**P.Earn	-	-	0.014	-0.020
			(0.83)	(-1.34)
*Endogenous**P.Acc	-	-	0.350	-0.228
			(0.87)	(-0.48)
*Endogenous**Late	-	-	-0.065	-0.052
			(-0.38)	(-0.48)
*Exogenous**Risk	-	-	-0.213	-0.219
			(-1.10)	(-1.44)
*Exogenous**P.Earn	-	-	0.028*	-0.008
			(2.16)	(-0.47)
*Exogenous**P.Acc	-	-	0.514[a]	-0.238
			(1.76)	(-0.54)
*Exogenous**Late	-	-	-0.211	-0.057
			(-0.98)	(-0.44)
*Framing**Risk	-	-	-0.151	-0.035
			(-0.82)	(-0.28)
*Framing**P.Earn	-	-	0.024	0.012
			(1.50)	(0.52)
*Framing**P.Acc	-	-	0.243	-0.117
			(0.65)	(-0.23)
*Framing**Late	-	-	-0.043	-0.103
			(-0.26)	(-1.10)
Constant	-0.257	1.835***	-0.258	1.835***
	(-0.35)	(4.24)	(-0.35)	(4.23)
N	6749	6749	6749	6749
Pseudo-R^2	0.109		0.113	

t statistics in parentheses
[a] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Below is the complete Table 6 including all variables. The table above the horizontal line corresponds to Table 6.

Table 18: Beliefs about the driving choices in the population

	Fast Belief (3a)	Slow Belief (3b)	Auto Belief (3c)	Fast Belief (4a)	Slow Belief (4b)	Auto Belief (4c)
AnyPolicy	-1.124	-1.127	2.251	-	-	-
	(-0.29)	(-0.32)	(0.88)			
AnyPolicy*Female	-2.851	-1.768	4.618*	-	-	-
	(-1.51)	(-1.19)	(2.19)			
Endogenous	-	-	-	-2.528	1.219	1.309
				(-0.59)	(0.32)	(0.49)
*Endogenous**Female	-	-	-	-4.995*	-0.827	5.821*
				(-2.49)	(-0.56)	(2.59)
Exogenous	-	-	-	3.286	-5.934[a]	2.648
				(0.72)	(-1.75)	(0.63)
*Exogenous**Female	-	-	-	-3.963	-0.715	4.678
				(-1.62)	(-0.30)	(1.37)
Framing	-	-	-	-1.991	0.128	1.863
				(-0.41)	(0.02)	(0.41)
*Framing**Female	-	-	-	0.309	-3.756*	3.447[a]
				(0.18)	(-2.42)	(1.79)
Female	1.395	2.848*	-4.243*	1.395	2.848*	-4.243*
	(0.86)	(2.73)	(-2.56)	(0.85)	(2.73)	(-2.56)
Risk	-0.0546	-0.102	0.157	-0.0548	-0.102	0.157
	(-0.09)	(-0.23)	(0.32)	(-0.09)	(-0.23)	(0.32)
P.Earn	0.258**	-0.124*	-0.134	0.258**	-0.124*	-0.134
	(3.28)	(-2.58)	(-1.30)	(3.29)	(-2.58)	(-1.30)
P.Acc	4.767*	-1.488	-3.279	4.773*	-1.495	-3.278
	(2.71)	(-1.95)	(-1.62)	(2.71)	(-1.96)	(-1.61)
Late	0.557	-5.316***	4.759**	0.557	-5.316***	4.759**
	(0.36)	(-7.60)	(3.17)	(0.36)	(-7.60)	(3.17)
AnyPolicy*Risk	0.861	0.539	-1.400[a]	-	-	-
	(1.22)	(0.90)	(-2.00)			
AnyPolicy*P.Earn	-0.078	0.066	0.013	-	-	-
	(-0.91)	(1.17)	(0.12)			
AnyPolicy*P.Acc	-1.732	2.168[a]	-0.436	-	-	-
	(-0.87)	(1.95)	(-0.19)			
AnyPolicy*Late	3.000	-0.255	-2.742	-	-	-
	(1.62)	(-0.22)	(-1.65)			
*Endogenous**Risk	-	-	-	1.097	0.257	-1.354[a]
				(1.54)	(0.44)	(-1.72)
*Endogenous**P.Earn	-	-	-	-0.125	0.121[a]	0.004
				(-1.52)	(1.91)	(0.04)
*Endogenous**P.Acc	-	-	-	-3.213[a]	2.798*	0.416
				(-1.73)	(2.44)	(0.19)
*Endogenous**Late	-	-	-	3.593	-0.888	-2.705
				(1.65)	(-0.59)	(-1.42)
*Exogenous**Risk	-	-	-	-0.053	1.434*	-1.381
				(-0.07)	(2.48)	(-1.60)
*Exogenous**P.Earn	-	-	-	0.005	0.048	-0.054
				(0.05)	(0.78)	(-0.45)
*Exogenous**P.Acc	-	-	-	1.373	1.566	-2.939
				(0.55)	(0.90)	(-0.95)
*Exogenous**Late	-	-	-	4.048	-1.481	-2.566
				(1.58)	(-0.85)	(-1.49)
*Framing**Risk	-	-	-	1.109	0.142	-1.251
				(1.28)	(0.16)	(-1.42)
*Framing**P.Earn	-	-	-	-0.092	0.001	0.091
				(-0.85)	(0.02)	(0.72)
*Framing**P.Acc	-	-	-	-2.888	1.714	1.173
				(-1.29)	(1.02)	(0.51)
*Framing**Late	-	-	-	1.386	1.442	-2.828
				(0.67)	(0.78)	(-1.29)
Constant	39.243***	25.688***	35.069***	39.238***	25.693***	35.068***
	(11.70)	(10.62)	(29.01)	(11.69)	(10.61)	(28.98)
N	6749	6749	6749	6749	6749	6749
R^2	0.049	0.059	0.015	0.074	0.078	0.056

t statistics in parentheses
[a] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Below is the complete Table 11 including all variables. The table above the horizontal line corresponds to Table 11.

Table 19: Driving choice (relative to Fast) - Response to first experienced fine

	Logit	*Multinomial Logit*	
	Non-Fast	Slow	Auto
	(5)	(6a)	(6b)
AfterFine	0.129	-0.728	0.670
	(0.14)	(-0.69)	(0.65)
AfterFine*Female	0.698[a]	0.649[a]	0.878[a]
	(1.94)	(1.66)	(1.73)
Female	-0.016	0.282	-0.391
	(-0.06)	(0.80)	(-1.16)
Risk	-0.087	-0.061	-0.120
	(-0.84)	(-0.55)	(-0.94)
P.Earn	-0.030***	-0.019*	-0.045***
	(-4.13)	(-2.35)	(-3.64)
P.Acc	-0.652***	-0.411[a]	-0.945***
	(-3.55)	(-1.89)	(-4.48)
Late	0.085	-0.181	0.385**
	(0.41)	(-0.58)	(2.70)
AfterFine*Risk	0.016	0.096	-0.039
	(0.14)	(0.84)	(-0.27)
AfterFine*P.Earn	-0.050*	-0.041[a]	-0.051[a]
	(-2.31)	(-1.70)	(-1.90)
AfterFine*P.Acc	-1.270**	-1.036*	-1.319*
	(-3.34)	(-2.11)	(-2.54)
AfterFine*Late	-0.053	0.260	-0.389[a]
	(-0.19)	(0.66)	(-1.75)
Constant	0.577	-0.418	0.224
	(1.20)	(-0.73)	(0.43)
N	2042	2042	2042
Pseudo-R^2	0.073	0.065	

t statistics in parentheses

[a] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix E - Contributing to punish Fast drivers

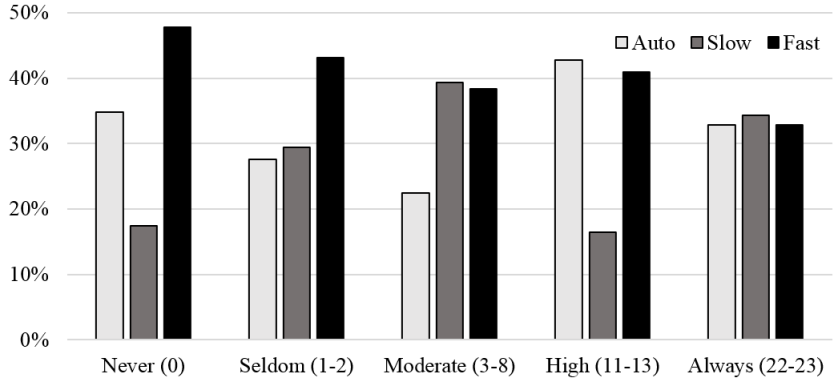
This section explores the factors that explain contribution decisions. Each of the 85 participants in *Endogenous* had the opportunity to contribute to the punishment fund in each round. As we mentioned before, participants who chose to contribute paid 1 which, in turn, would increase price of the fine paid by Fast drivers by 2.5. Out of a total of 1,854 opportunities to contribute, 209 contributions are made (11.3%). Slow drivers contribute 13.9% of the time, which is higher than Fast drivers or Auto drivers (10.4% and 10.5%, respectively). Table 20 groups the participants in *Endogenous* into contribution "types" based on the number of times they contributed. 48.3% of participants contributed to the fund at least once whereas three participants contributed in every round.[17]

Table 20: Number of participants within each Contribution Type

Type	Never (0)	Seldom (1-2)	Moderate (3-8)	High (11-13)	Always (22-23)
# of participants	44	23	10	5	3
	(51.7%)	(27.1%)	(11.8%)	(5.9%)	(3.5%)

Figure 7 displays the percentage of each driving choice separated by contribution types. Higher contribution types do not monotonically relate to any driving choice (which is also true when ignoring the "Always" contribution type).

Figure 7: Participants choices by punishment types



Female participants contribute to the punishment fund 13.5% of the time, which is higher than male participants (8.6%). However, these differences are based on only 8 sessions of

[17]Interestingly, the three "Always" participants employed the same driving choice in every round they played. One participant chose Fast every round, one participant chose Slow every round, and one participant chose Auto every round.

Endogenous and a logit model which clusters at the session level shows this difference to be insignificant.

Table 21: Contribution choice

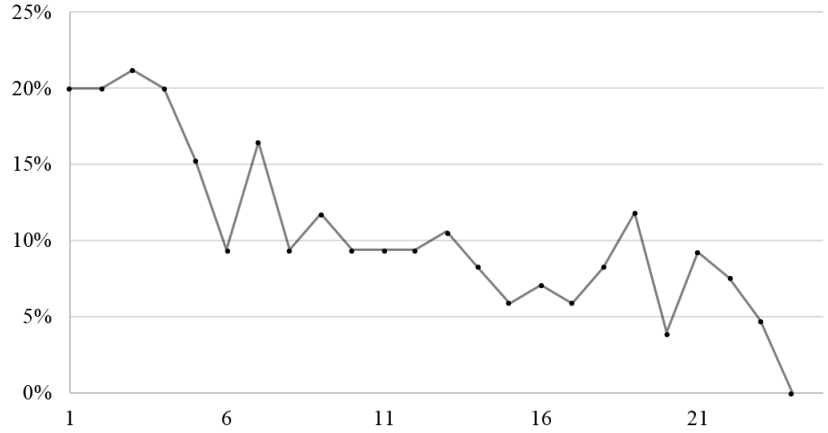
Female	0.474
	(0.83)
Risk	-0.113
	(-0.53)
P.Earn	-0.016
	(-1.14)
P.Acc	-0.456
	(-1.51)
Late	-0.726***
	(-3.72)
N	1854
Pseudo-R^2	0.037

t statistics in parentheses

[a] $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Risk preference and outcomes from the previous round are not associated with contribution decisions. The only significant predictor of contribution decisions is whether the participant is in the late rounds of the experiment. Figure 8 shows that the proportion of participants contributing to the punishment fund declines over the duration of the experiment.

Figure 8: Contribution decisions by round

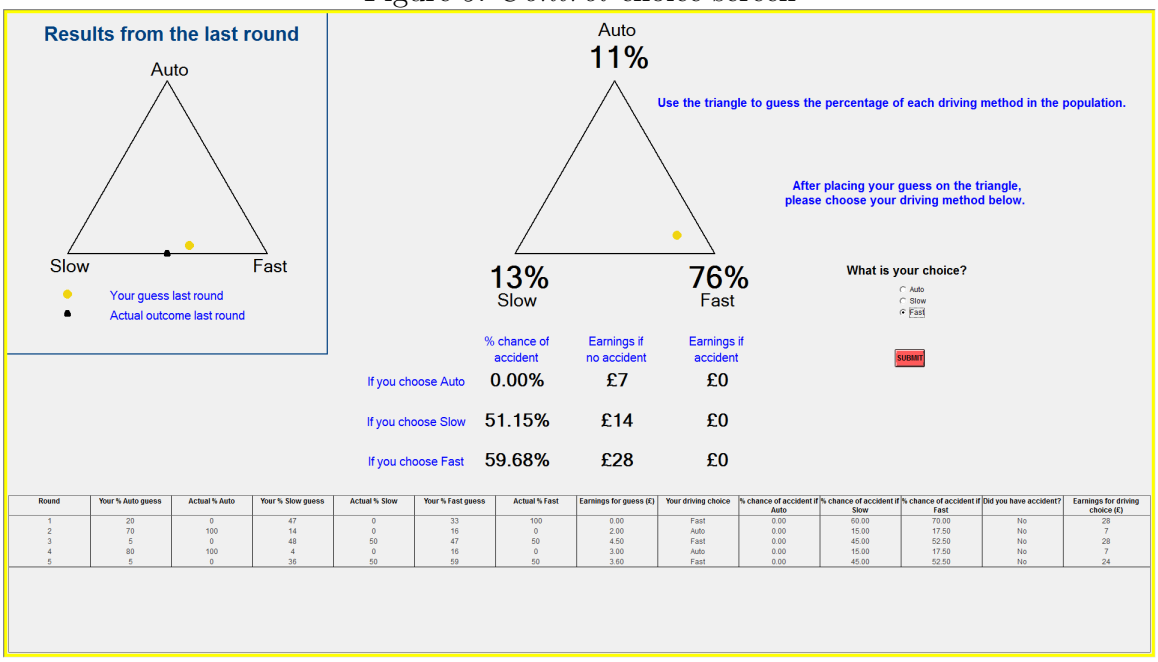


Declining punishment enforcement is famously observed in prisoner's dilemma ([13]) and

public goods games ([25]) both which resemble the punishment decision in *Endogenous*. With this one exception, our data suggest that the decision to contribute to the punishment fund is orthogonal to driving choice, gender, risk preference, and outcomes from the previous round.

Appendix F - Screen shots

Figure 9: *Control* choice screen



Round	Your % Auto guess	Actual % Auto	Your % Slow guess	Actual % Slow	Your % Fast guess	Actual % Fast	Earnings for guess (£)	Your driving choice	% chance of accident if Auto	% chance of accident if Slow	% chance of accident if Fast	Did you have accident?	Earnings for driving choice (£)
1	20	0	47	0	33	100	0.00	Fast	0.00	60.00	70.00	No	28
2	70	100	14	0	16	0	2.00	Auto	0.00	15.00	17.50	No	7
3	5	0	48	50	47	50	4.50	Fast	0.00	45.00	52.50	No	28
4	80	100	4	0	16	0	3.00	Auto	0.00	15.00	17.50	No	7
5	5	0	36	50	59	50	3.60	Fast	0.00	45.00	52.50	No	24

Figure 10: *Framing* choice screen

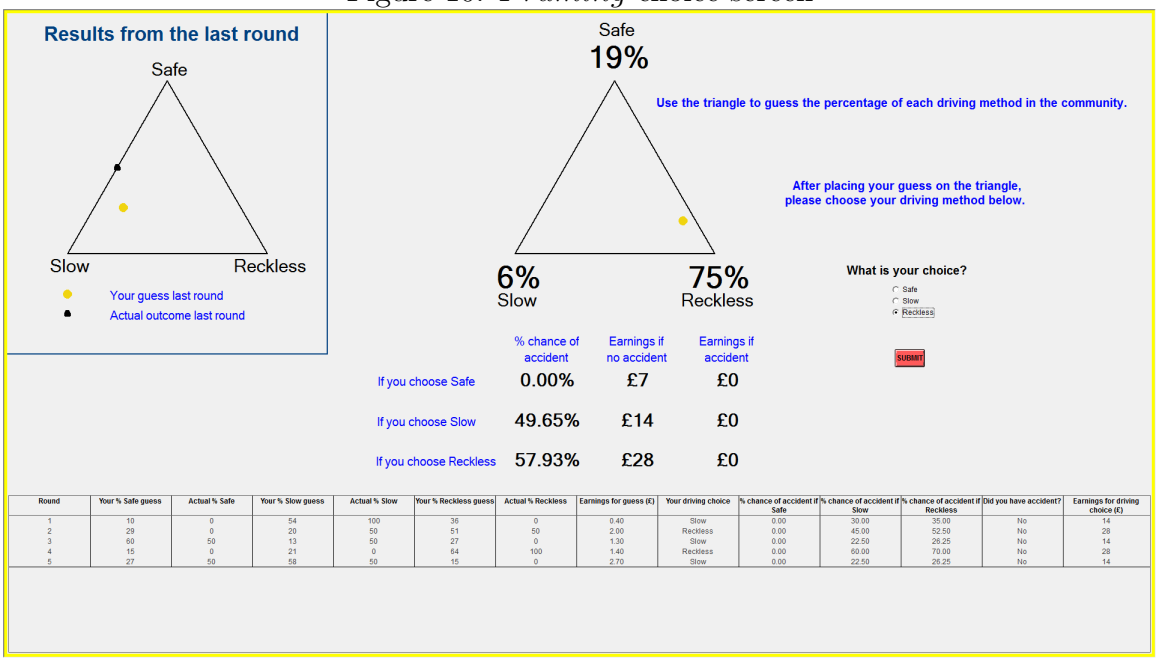


Figure 11: *Exogenous* punishment choice screen

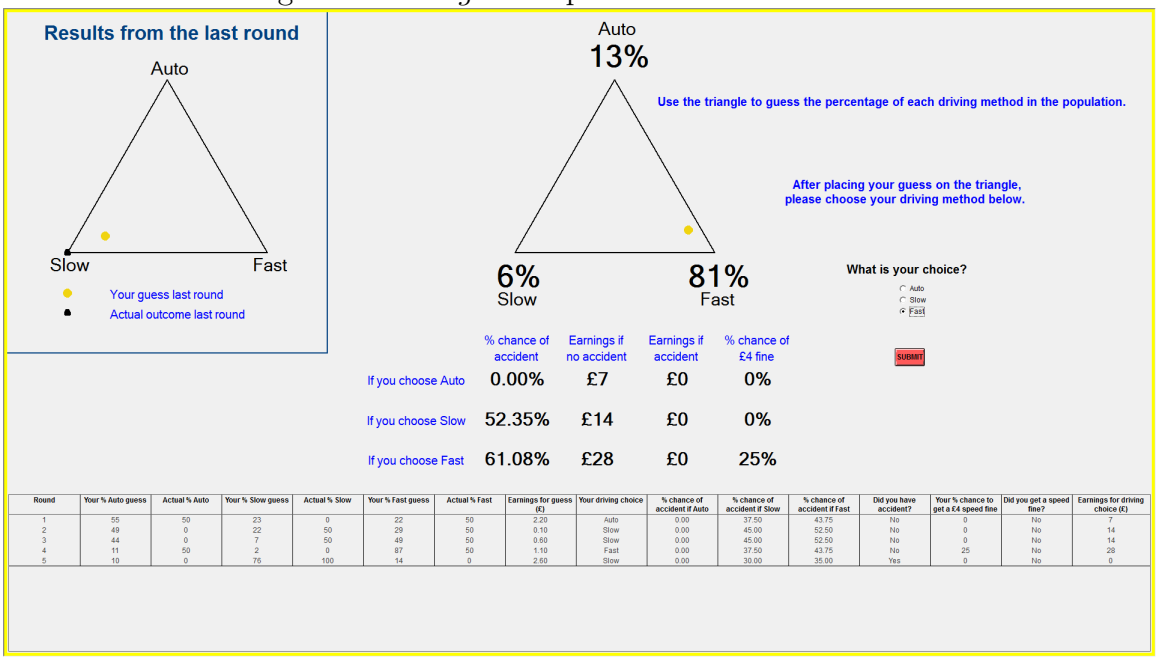
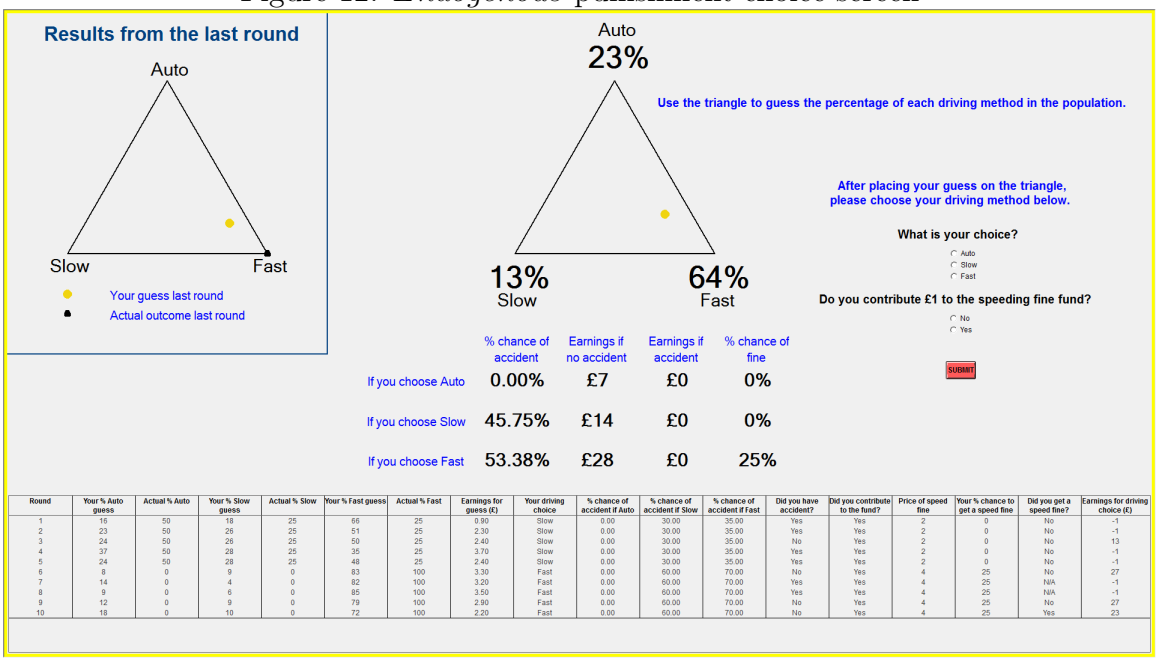


Figure 12: *Endogenous* punishment choice screen



Round	Your % Auto guess	Actual % Auto	Your % Slow guess	Actual % Slow	Your % Fast guess	Actual % Fast	Earnings for guess (£)	Your driving choice	% chance of accident if Auto	% chance of accident if Slow	% chance of accident if Fast	Did you have accident?	Did you contribute to the fund?	Price of speed fine	Your % chance to get a speed fine	Did you get a speed fine?	Earnings for driving choice (£)
1	16	50	18	25	66	25	0.90	Slow	0.00	30.00	35.00	Yes	Yes	2	0	No	-1
2	23	50	26	25	51	25	2.30	Slow	0.00	30.00	35.00	Yes	Yes	2	0	No	-1
3	24	50	28	25	50	25	2.40	Slow	0.00	30.00	35.00	No	Yes	2	0	No	13
4	37	50	28	25	35	25	3.70	Slow	0.00	30.00	35.00	Yes	Yes	2	0	No	-1
5	24	50	28	25	48	25	2.40	Slow	0.00	30.00	35.00	Yes	Yes	2	0	No	-1
6	8	0	9	0	83	100	3.30	Fast	0.00	60.00	70.00	No	Yes	4	25	No	27
7	14	0	4	0	82	100	3.20	Fast	0.00	60.00	70.00	Yes	Yes	4	25	NA	-1
8	9	0	6	0	85	100	3.50	Fast	0.00	60.00	70.00	Yes	Yes	4	25	NA	-1
9	12	0	9	0	79	100	2.90	Fast	0.00	60.00	70.00	No	Yes	4	25	No	27
10	18	0	10	0	72	100	2.20	Fast	0.00	60.00	70.00	No	Yes	4	25	Yes	23

Figure 13: *Control* results screen

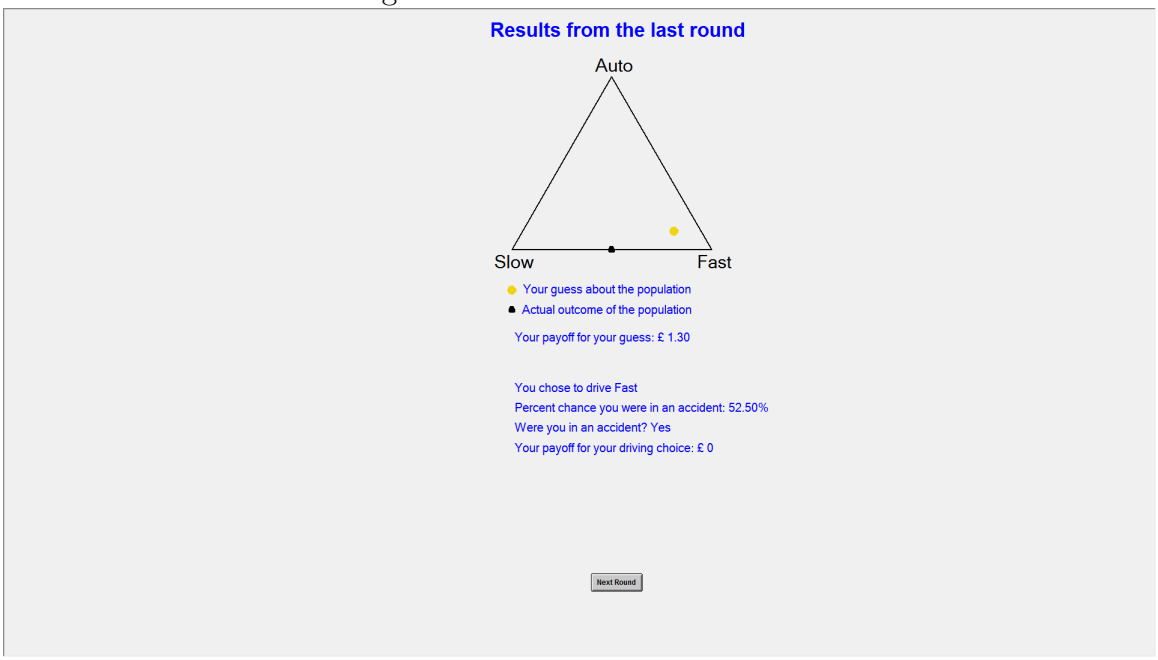


Figure 14: *Framing* results screen

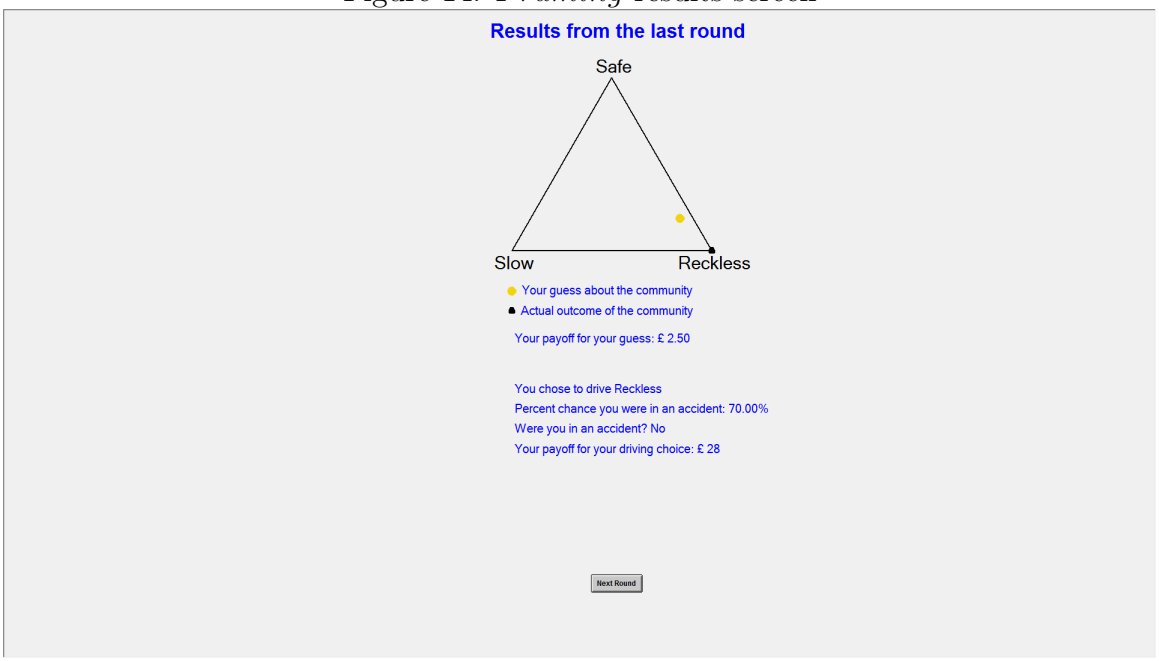


Figure 15: *Exogenous* punishment results screen

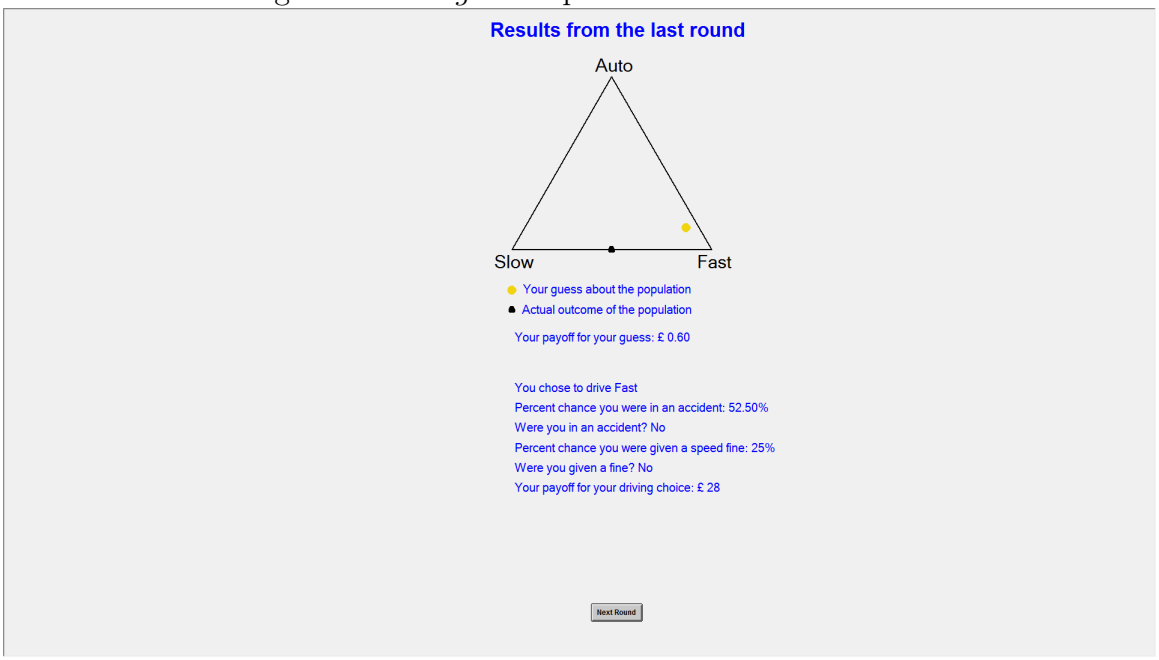


Figure 16: *Endogenous* punishment results screen