

Willingness-to-pay for Warnings

A. Gaduh, P. McGee and A. Ugarov

September 7, 2023

- Many real-life situations require alarm designers to determine risk thresholds for sounding alarms: medical tests, fire alarms, extreme weather warnings, etc
- This requires balancing the costs of false-positive and false-negative events
- There is little theory and even less empirical studies on people preferences over signal characteristics

- ① How do signal characteristics affect protective response?
- ② Which signal characteristics people prefer? Two hypotheses:
 - Risk-neutral decision model provides a good description of user's preferences
 - Subjects put equal weights on costs coming from false-positive and false-negative events

Findings

- There are significant deviations from the risk-neutral model both for protective decisions and for willingness-to-pay for signals
- WTP has excess sensitivity to false-negative rates for low priors, and lower sensitivity for high priors (vice versa for false-positive)
- Subjects tend to have excessive reactions to false-positive rates with negative signals and excessive reactions to false-negative rates for positive signals (alarms)
- This effect partially comes from distorted beliefs on posterior chances of false-positive/false-negative event conditional on a signal
- This pattern is not consistent with EU framework, but most consistent with decision-making heuristic in which subjects do not differentiate between false-positive and false-negative rates when choosing signals

Overview of the Experiment

Willingness to pay (WTP) for signals

- An insurance experiment:
 - Two states of the world: bad ($\omega = 1$) and good ($\omega = 0$)
 - Probability of a bad state is $P(\omega = 1) = \pi$
 - Bad state \implies loss of $\$L$
 - A perfectly protective insurance can be purchased for $\$c$
- Subject can purchase a signal s before purchasing the insurance:
 - A signal is characterized by its true-positive ($P(s = 1|\omega = 1)$) and true-negative rates ($P(s = 0|\omega = 0)$)

WTP for Signals

Theory

- Theoretically, what should be the WTP for a signal?
- If bad states are a priori rare ($\pi L \ll c$) \implies never protect without a signal
- The theoretical WTP b for an expected utility maximizer given a signal s is a solution b^* to the following:

$$\begin{aligned} P(s=1)u(Y_0 - b^* - c) + \pi P(s=0|\omega=1)u(Y_0 - b^* - L) + \\ + (1 - \pi)P(s=0|\omega=0)u(Y_0 - b^*) = \\ = (1 - \pi)u(Y_0) + \pi u(Y_0 - L) \end{aligned}$$





- A risk-neutral agent would therefore pay:

$$b^* = \pi(1 - P(s=0|\omega=1))L - P(s=1)c$$

- ① Box with 20 white and black balls (black ball=bad state)
- ② Assumptions:
 - Protection cost is \$5
 - Loss without protection is \$20
 - Cost-loss ratio is $c/L = 5/20 = 0.25$
- ③ Signal is an unreliable hint about the ball color
- ④ Vary the prior probability of bad state and the signal's information structure

Representing Signals

- A subject receives a noisy signal as a hint from one of the gremlins:

Ball/Gremlin	<i>Honest gremlin:</i> 	<i>White-swamp gremlin:</i> 	<i>Black-swamp gremlin:</i> 
	The Ball is white!	The Ball is white!	The Ball is black!
	The Ball is black!	The Ball is white!	The Ball is black!

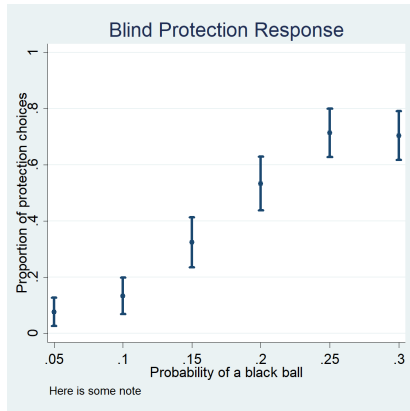
- 1 **Blind Protection Game:** Protection response conditional on prior probability
- 2 **Informed Protection Game:** Protection response conditional on prior probability and signal
- 3 **Belief Elicitation:** Subjects beliefs about the bad state's probability conditional on prior and signal
- 4 **WTP Elicitation:** Willingness-to-pay for each signal

Going over all the treatments

- Out of 4 tasks, two are novel (informed protection and WTP elicitation for a signal)
- We report basic results to show that the subjects' behavior is largely sensible in all the tasks
- Responses demonstrate known biases in existing tasks

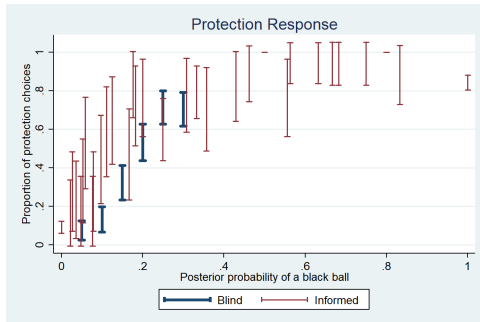
Blind Protection (No Signal)

- There is more protection when the probability of black ball is higher
- Responses are mostly monotonic in probability
- There are both risk-averse and risk-loving subjects



Informed Protection

- Protection rates increase with the posterior probability of a bad event (black ball)
- Average protection probability is no longer monotonic in posteriors due to lower N of obs per prob and due to posterior estimation mistakes



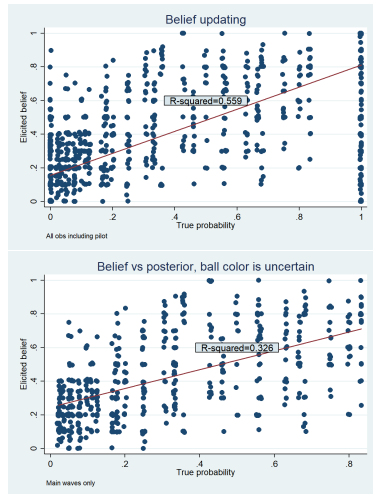
IP Big Picture

- Overprotect in response to white signals, underprotect in response to black signals without false-positive: can be explained by risk preferences
- Overprotect in response to white signals with FP (cannot be explained by risk preferences)!

Signal	False-pos.	False-neg.	% protect	Posterior	Optimal	P(=optimal)
White	No	No	0.000	0.067	0.000	0.000
White	No	Yes	0.100	0.333	0.000	0.000
White	Yes	No	0.000	0.130	0.000	0.000
White	Yes	Yes	0.131	0.564	0.121	0.000
Black	No	No	1.000	0.846	1.000	0.000
Black	No	Yes	1.000	0.841	1.000	0.000
Black	Yes	No	0.550	0.833	0.870	0.355
Black	Yes	Yes	0.483	0.886	0.871	0.685

Belief Elicitation

- Correlation between beliefs and posteriors
- Large dispersion of errors
- Includes surprisingly large errors even for certain signals

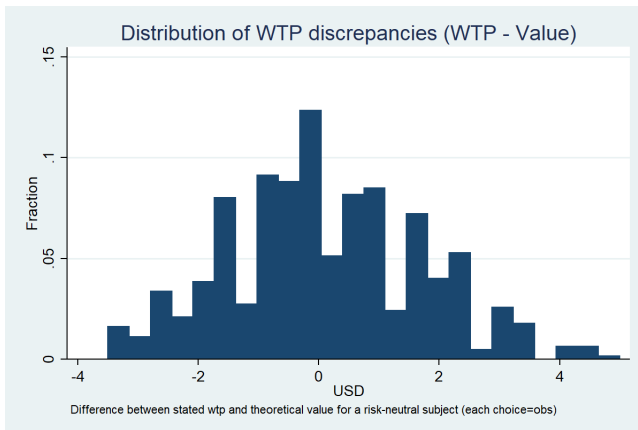


- Overestimating chances of negative events for white signals, including the cases without FN
- Underestimating chances of negative events for black signals both for honest signals and when only FN events are possible

Table: Average Belief Error by Signal Type

	False-pos.	False-neg.	Signal	Posterior	Belief error	P(= 0)
(1)	No	No	White	0.000	0.050	0.000
(2)	No	Yes	White	0.100	0.122	0.000
(3)	Yes	No	White	0.000	0.122	0.000
(4)	Yes	Yes	White	0.131	0.218	0.000
(5)	No	No	Black	1.000	-0.163	0.000
(6)	No	Yes	Black	1.000	-0.279	0.000
(7)	Yes	No	Black	0.550	0.039	0.130
(8)	Yes	Yes	Black	0.483	0.048	0.021

Distribution of WTP Discrepancies



WTP for Signals: Big Picture

- Overpaying for signals when both false-positive and false-negative events are possible
- Small and insignificant discrepancies for other signal types

False-positive	False-negative	Mean WTP discrepancy	P(= 0)
No	No	-0.106	0.433
No	Yes	0.143	0.250
Yes	No	0.081	0.502
Yes	Yes	0.492	0.000

- Next, we test main hypotheses of this study
- Calculate the difference between reported WTP and theoretical WTP
- Regress the difference on FP costs $((1 - p)P(s = 1|\omega = 0)c)$ and FN costs $(pP(s = 0|\omega = 1)L)$
- Coefficients should be zero if the theoretical model is correct

WTP for Signals: Determinants

Table 3 from the latest paper draft goes here

- No evidence of relative underweighting of FP or FN signals on average ($\beta_{FP} = \beta_{FN}$)
- Our results exhibit the following anomaly:
 - ① Excess sensitivity to false-negative costs for low priors; lack of sensitivity for high priors
 - ② Underresponse to false-positive costs for low priors; excess sensitivity for high priors
- Risk preferences does not explain FP/FN sensitivity and at best only partially explains the pattern of changing sensitivity with priors
- Next, we explore potential explanations for the anomaly

Anomaly Reframed

- Note that the apparent heterogeneity comes from contrasting with the theoretical value! The sensitivity of WTP to FP/FN rates shows little variation by prior.
- The coefficients on FP/FN rates are also surprisingly close and their differences are statistically insignificant
- \implies Suggests that subjects poorly differentiate between FP and FN rates when choosing WTP

	(1)	(2)	(3)	(4)
	0.1	0.2	0.3	0.5
FP rate	-2.91*** (1.1)	-2.08** (1.0)	-4.46*** (1.0)	-3.25** (1.3)
FN rate	-2.48** (1.1)	-2.73*** (1.0)	-3.7*** (1.0)	-3.65*** (1.3)
P(FP rate=FN rate)	.792	.669	.617	.832
Adjusted R^2
Observations	159	153	159	153

Standard errors in parentheses

Tobit regression of WTP, constant omitted

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Potential Explanation for the Anomaly

- ① If subjects do not differentiate between FP and FN rates when choosing WTP that would explain both the anomaly and will be consistent with FP/FN coefficients being equal to each other:
 - Priors have opposite effects on the frequency of FP/FN events and these effects partially cancel each other if we treat them as the same thing
- ② Several subjects report this verbal explanation for IP choices post the experiment: e.g. *"I trusted honest golems fully, and did not put much stock in the swamp golems."*; *"If the hint was from one of the honest gremlins then I didn't choose to protect because they could only tell the truth. If there were any just black or just white gremlins then I decided to protect because the information they give isn't helpful"*
- ③ If this explanation is true, then we should expect that subjects in other treatments also react to FP/FN rates in ways consistent with this heuristic

Belief Biases by FP/FN rates

- If subjects confuse FP/FN rates, then FP rates would increase beliefs when the signal is white and FN rates would decrease beliefs when the signal is black
- Estimation results for belief error indeed demonstrate this pattern

Table: Belief Elicitation: When Mistakes Happen

	(1) All	(2) S=White	(3) S=Black
FP rate	.6*** (0.1)	.292*** (0.1)	.908*** (0.1)
FN rate	.0108 (0.1)	.273*** (0.1)	-.251*** (0.1)
Constant	-.0784*** (0.0)	.314*** (0.0)	-.47*** (0.0)
Subject FE	Yes	Yes	Yes
Observations	1248	624	624
Adjusted R^2	0.15	0.41	0.52

Standard errors in parentheses

Dep. variable: reported belief - posterior probability

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Informed Protection Biases by FP/FN rates

- If subjects confuse FP and FN rates, we should also expect excessive positive reaction to FP rates for white signals and excessive negative sensitivity to FN rates for black signals
- Note: excessive reaction here is any reaction to FP/FN rates happening conditional on posteriors.
- We find the positive significant coefficient for FP rates for white signals, but the FN rate coeff for black signals is insignificant (though negative as predicted)
- Controlling for reported beliefs (which are already erroneous in the same direction) reduces the magnitude of excess FP rate sensitivity but it still remains significant: either the bias is not completely due to beliefs or beliefs drift between treatment

IP Biases by FP/FN rates: Results

	(1)	(2)	(3)	(4)
FP rate x (S=White)	.434*** (3.2)	.508*** (3.4)	.235* (1.8)	.293** (2.0)
FN rate x (S=White)	.43* (1.8)	.412 (1.7)	.0513 (0.2)	.0366 (0.1)
FP rate x (S=Black)	-.128 (-0.3)	-.0328 (-0.1)	-.187 (-0.4)	-.114 (-0.2)
FN rate x (S=Black)	.0434 (0.4)	-.067 (-0.4)	.000394 (0.0)	-.0827 (-0.5)
S=Black	.342* (2.0)	.382* (1.8)	.196 (1.1)	.225 (1.1)
p>0.2	.0504** (2.5)	.0434* (1.9)	.0299 (1.5)	.0251 (1.1)
FP rate x (p>0.2)		-.131 (-0.8)		-.101 (-0.6)
FN rate x (p>0.2)		.221 (1.3)		.165 (1.1)
Subject FE	Yes	Yes	Yes	Yes
Posterior	Yes	Yes	Yes	Yes
Beliefs	No	No	Yes	Yes
Observations	1248	1248	1248	1248
Adjusted R^2	0.53	0.53	0.55	0.55

t statistics in parentheses

* $p < 0.10$. ** $p < 0.05$. *** $p < 0.01$

Taking Stock on Explaining the Anomaly

- Risk preferences do not seem to explain the anomaly
- Poor differentiation between FP and FN rates is obvious in the WTP task (coeffs are equal for each prior) and can completely explain the anomaly
- Choices in BE and IP task exhibit patterns consistent with this explanation
- Tailored experimental tests are needed to confirm that belief biases are indeed responsible for the anomaly

Alternative Explanations

- At the end, we agreed to go over other less likely explanations which can pop up in discussions:
 - Neglecting the frequency of FP/FN events
 - Value of non-instrumental information
 - Anchoring (not added here)

Do Subjects Neglect to Account for Frequency of FP/FN events?

- The anomaly is also consistent with subject neglecting prior probabilities when evaluating the effects of FP and FN rates:
 - For the same FN rate $P(s = 0|\omega = 1)$, FN event is more likely when the priors are high as its probability is $pP(s = 0|\omega = 1)$ (the reverse is true for FP)
 - In contrast, experimental subjects reduce WTP with FP/FN rates but do it uniformly for all the priors
 - Priors still affect demand but only directly
- This is a reasonable heuristic to evaluate the value of information: account for priors, and correct according to false-positive and false-negative rates (these variables are immediately given and do not require extra computation)

Neglecting Frequency of FP/FN events

- We can test this explanation using WTP data in two ways:
 - Rewrite the interaction as the base rate plus interaction and test the joint significance of the interaction terms
 - Test for joint significance of the full set of flexible interactions between prior levels and FP/FN rates
- The first test cannot reject the null hypothesis, but the interaction terms are jointly significant (with or without FE) at 5%
- The change in explanatory power is very small and coefficient magnitudes are non-sensical (smaller for 0.3 rather than for 0.5)

Neglecting Frequency of FP/FN events

	(1)	(2)	(3)	(4) FE
Prior_change \times FP rate	-.392 (0.5)		-.457 (0.4)	
Prior_change \times FN rate	-.0869 (0.1)		-.08 (0.1)	
plevel=200 \times FP rate		.0497 (0.2)		.0668 (0.2)
plevel=300 \times FP rate		-.281*** (0.1)		-.281*** (0.1)
plevel=500 \times FP rate		-.103 (0.2)		-.0857 (0.2)
plevel=200 \times FN rate		-.000453 (0.0)		.000683 (0.0)
plevel=300 \times FN rate		-.0384* (0.0)		-.0384* (0.0)
plevel=500 \times FN rate		-.0295 (0.0)		-.0284 (0.0)
P(beta_X=0)	.544	.0286	.53	.0276
Adjusted R^2	.158	.156	.543	.543
Observations	624	624	624	624

Standard errors in parentheses

Controlling for FP/FN rates, prior, constant omitted

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Alternative Explanations: Paying for non-instrumental information

- Humans often put positive value on information not changing their decisions (citations)
- Signals with zero theoretical value can have positive elicited WTP
- We see subjects paying positive amounts for signals with zero theoretical value as well as paying for signals not affecting their decisions in IP task, but it is not clear if those are reasoning mistakes or genuine preferences
- Explaining the pattern would require non-instrumental part of value to have positive sensitivity to FP/FN rates for some priors and negative for others
- A priori there is no reason for the non-instrumental value to **decrease** with either FP or FN costs

Potential Further Tests

- Extra treatment: provide natural frequencies of false-positive/false-negative events before eliciting WTP
 - Example: *"out of 1000 experimental runs, roughly 250 involve a gremlin saying "Black" while the ball is white"*
- Elicit the probability of a black signal $P(S = B)$ within the belief elicitation task to see if beliefs on FP/FN events signals are biased
- Practice rounds so that subjects learn the frequencies of FP/FN events for each signal