July, 2025 ERI-CES DP 25/03

# Exploring the Privacy Paradox: An Experimental Investigation of Privacy-Preserving behavioural Responses in Online Shopping

> Penélope Hernández
Universitat de València and ERI-CES, Spain

> Antonio J. Morales
Universidad de Málaga, Spain

> Zvika Neeman
Tel Aviv University, Israel

> Jose M. Pavía
Universitat de València, Spain

www.erices.es

Discussion Papers
in Economic Behaviour

Universitat de València

# Exploring the Privacy Paradox: An Experimental Investigation of Privacy-Preserving behavioural Responses in Online Shopping*

Penélope Hernández, Antonio J Morales, Zvika Neeman, and Jose M. Pavía

July 28, 2025

**Abstract**

This paper utilises (an observational approach within) a controlled laboratory experiment to investigate the nature of the privacy paradox. Participants are assigned a type exogenously and engage in online shopping to earn monetary rewards, while their shopping behavior is observed by an AI that aims to learn their type. Our findings indicate that participants willingly disclose significant amounts of private information, and persist in doing so even after receiving explicit information regarding the AI's ability to learn about their type. However, we observe that the adoption of two mechanisms, namely "explainable AI" and a "privacy APP", leads participants to adopt privacy-preserving shopping habits. Notably, this change in behavior occurs even in scenarios where the disclosure of private information has no impact on the monetary rewards. Our findings suggest that a plausible reason individuals share extensive personal information online stems from their lack of access to technologies enabling them to engage online while safeguarding their privacy.

**JEL CODES:** C91, D02, D90.

**KEYWORDS: Intrinsic preference for privacy, experiment, privacy paradox, explainable AI, privacy-APP**

*The authors acknowledge the support of the Consejería de Economía, Innovación, Ciencia y Empleo, Junta de Andalucía, grant PY20-00069. Hernández and Pavía acknowledge the support of the Spanish Ministry of Science and Innovation, grant PID2021-128228NB-I00, and by Consellería d´ Innovació, Universitats, Ciència i Societat Digital, Generalitat Valenciana, grant AICO/2021/257.

# 1 Introduction

A concern for privacy is defined as a desire to keep certain aspects of an individual's personal life inaccessible to others ([35], [36], [5]). On the one hand, surveys systematically show that people are highly concerned about privacy ([33],[10], [23], [11]). But, on the other hand, people often behave in ways that reveal significant amounts of personal information (such as contact details, financial information, or browsing history) to others online. This contradiction is known in the literature as the "Privacy Paradox" ([8], [7], [25], [15]).

There are several alternative explanations for why people may end up revealing personal information, which can be classified based on their level of awareness. Some individuals may not fully understand the situation or its implications for privacy (confusion). Others, while not confused, may fail to realize how much information they are disclosing (ignorance or mistaken beliefs). Finally, some people share personal data because they believe they have no other choice (resigned privacy).

Analyzing online user behavior in real-world settings makes it difficult to determine the nature of the so-called "privacy paradox." We design a payoff-neutral lab environment in which to isolate intrinsic privacy preferences, we test whether participants keep disclosing even after they learn that privacy disclosure does not help or hurt them monetarily, and we examine interventions.

In the laboratory, participants, endowed with a private type at the beginning of the experiment, make repeated shopping decisions that allow them to accumulate monetary rewards in an AI-driven online platform. Participants' purchasing decisions inform about their types and the artificial intelligence learns about the private type of participants.

Using lab data, we gain deeper insights into the nature of privacy behavior because we can (i) control for confusion, (ii) measure users' beliefs about the amount of information they disclose, and (iii) provide participants with tools to engage online while safeguarding their privacy. Our results suggest that a plausible reason individuals share extensive personal information online stems from their lack of access to technologies that enable them to participate online while protecting their privacy. We successfully test two mechanisms that would allow individuals to better align their behavior and their preferences regarding privacy: "explainable AI" and "privacy APP", where APP stands from Application for Privacy Protection.

As in any real-life setting, different shopping behaviors will induce different learning processes in artificial intelligence, resulting in varying levels of private information shared by consumers. Because the AI-driven platform -using the information gathered by the artificial intelligence- selects which products to offer and at what prices, participants might wonder how the private information revealed by their shopping behavior will be used by firms to select products and prices and therefore how it will affect their monetary payoffs arising from their transactions.

This instrumental component of the preference for privacy was the way in which the preferences for privacy first entered into the economics analysis at the end of the twentieth century. It was the natural way because it fitted nicely into the framework of the economics of information, whose focus was on how the existence of asymmetric information between the two parts of an economic transaction impacted the terms of the transaction. Implicit in this approach is that privacy is not valued per se but that it confers some economic advantage. For example, [4] or more recently [17] consider the rational behavior on part of the consumers to not revealing their private preferences to avoid price discrimination.

This economics approach relied on the existence of stable and rational preferences for privacy which could in principle, be empirically elicited by asking customeinrs about their willingness to pay or their willingness to accept (*WTP* and *WTA*). But, the empirical literature on privacy elicitation soon discovered that the elicited values of privacy were of a subjective nature and context-dependent (see [3] for a comprehensive survey). Moreover, the picture of the empirical analysis of privacy preferences got complicated when attending the philosophy and legal literature, that acknowledge that individuals regard privacy with the same importance as other inherent values, such as freedom and autonomy. Hence, there might also be an intrinsic component to privacy.

To disentangle the instrumental component of the preference for privacy from the intrinsic component that captures the participant's inherent preference for privacy -using the terminology in the dual-privacy theory proposed by [21]- we fine-tune our shopping environment to make (expected) monetary rewards accumulated by participants invariant to their amount of private information revealed by their shopping behavior. Because participants are informed about this feature of the experimental design, our experimental results will neatly inform about the existence and magnitude of their intrinsic preferences for privacy even in a setting as artificial as ours.

The key feature of our experimental design is the payoff-neutral shopping environment. We next describe it. Each participant interacts with one firm (individual decision making). At each round, the firm selects which of two products to offer, $A$ or $B$, and at what price, high or low, and the consumer decides whether to buy or not.

At the beginning of the experiment, each participant is randomly assigned one of two types, $A$-dominant or $B$-dominant. The participant's type is held constant throughout the experiment. At each round, the participant is assigned temporal preferences, that apply only in that round. With probability 0.75, $A(B)$-dominant types have a preference for product $a(b)$, and with the complementary probability 0.25, $A(B)$-dominant types have a preference for product $b(a)$. Whenever the participant's type coincides with her temporal preferences (i.e, an $A(B)$-dominant type has a temporal preference for product $a(b)$), then with probability 0.75 the participant has a higher willingness to pay (WTP) for that product, and with probability 0.25 a lower willingness to pay for it. Whenever the participant's type does not coincide with her temporal preferences, then with probability 0.75 the participant has a lower willingness to pay for that product, and with probability 0.25 a higher willingness to pay for it.

In each round, the participant's payoff is given by the difference between her WTP for the product that is offered to her and the price paid. The WTP for a product that does not match the participant's temporal preferences is zero. It follows that the participant should rationally refrain from purchasing a product that does not match her temporal type. Because the participant's temporal preferences are correlated with her type, the participant's purchasing decisions inform about her type.

In order to achieve a payoff-neutral shopping environment, we set the participants' WTP values and the prices in the following way. Let $p_A^t$ denote the posterior probability that the participant is $A$-dominant in round $t$, based on her shopping behavior. The experiment is set up so that in each round, the firm offers product $A$ if $p_A^t > 0.5$, and product $B$ if $p_A^t < 0.5$ (and randomizes if $p_A^t = 0.5$). Moreover, when the firm offers product $A(B)$, then it sets a high price with probability $p_A^t$ $(1 - p_A^t)$, and a low price with the complimentary probability $1 - p_A^t$ $(p_A^t)$.[1]

Consider the following two purchasing strategies for the participant: (i) purchase whenever the quoted price is less than or equal to the WTP, and (ii) purchase only when you stand to make the largest possible profit. Namely, purchase only when the participant has a high WTP for the product and the price is low. Each of these strategies induces a different process of learning about the participant's type. We calibrate the participant's WTP values and the prices so that the participant's payoff would be equal under these two strategies, taking the appropriate learning process into account. In the experiment, the firm behaves as described above, but participants are, of course, free to choose any purchasing strategy they want. Nevertheless, the participants' payoffs are seen to be similar to each other, as explained below.

We conduct four treatments: Control, Awareness, Info (Explainable AI), and Privacy App. The distinctions among these treatments are based on the information provided after Block 1. Participants will then engage in a new block of 30 rounds, during which they will make more purchasing decisions in the same platform. In the Control treatment, participants are informed that their shopping behavior is monitored by an AI system capable of inferring their experimentally induced preferences. In addition, participants in the other treatments are asked, prior to their second shopping experience, to estimate how much private information they believe they are disclosing. The Awareness treatment simply makes participants aware of this fact. In the Info treatment, participants receive a comprehensive explanation of how the AI learns about their private information from their shopping habits. The Privacy App treatment allows par-

[1] The belief $p_A^t$ was updated in the following way. The distribution of the firm's belief that the consumer is $A$-dominant was initially given by a Beta distribution with parameters $\alpha = \beta = 1$, and the firm's posterior belief was initially set to be equal to the expectation of this distribution, $p_A^1 = \frac{1}{2}$. Then, any purchase of product $A$ increased the parameter $\alpha$ by one, and any purchase of product $B$ increased the parameter $\beta$ by one. At any round $t$, the firm's posterior belief $p_A^t$ was given by the expectation of the updated posterior Beta distribution.

ticipants to control the amount of private information they disclose through a ”privacy APP,” which is designed to help them maintain their privacy without compromising their monetary rewards.

Our experiment finds that, even in the somewhat artificial setting of a laboratory, we observe behaviors that suggest an intrinsic preference for privacy. When participants were informed about how the AI operates, they adjust their shopping behavior to minimize the amount of private information disclosed, even knowing that this revelation would not affect their monetary payoffs. Also, when offered the Privacy APP, one-third of the participants committed to using it.

By analyzing data from both interventions, we can identify certain characteristics of “privacy lovers” individuals who actively protect their private information when given the means to do so. We find that these privacy lovers often perceive themselves as disclosing large amounts of private information, and they tend to be female and risk-averse, regardless of whether they choose to disclose a significant amount of information or remain more reserved.

Regarding those who use the Privacy APP, two variables associated with the Big Five personality traits—agreeableness and conscientiousness—showed an negative relationship with APP adoption. It could be argued that individuals who are more self-aware may not feel the need for the additional assistance that the app provides. However, for those who find managing stress challenging, using the APP could be beneficial, which explains the positive correlation with the neuroticism trait. Risk aversion significantly influences this behavior; individuals who are more cautious about taking risks tend to be less inclined to adopt the APP. One interpretation of this behavior is that risk-averse individuals may hesitate to rely on a machine for recommendations, particularly if they believe it could lead to a loss of control.

## Related Literature

A substantial body of literature explores the subject of privacy, providing a thorough examination from various angles, including environmental psychology, economics, consumer behavior, among others. From the perspective of environmental psychology, Altman’s seminal work, [5], posits that privacy is not a static state but rather a regulatory process through which individuals and groups manage their interactions with others. Additionally, [4] investigate how companies utilize purchase history to influence pricing strategies, raising critical questions about the implications for consumer privacy. They address the potential advantages of personalized pricing, such as enhanced efficiency and increased consumer surplus, alongside the associated risks, including discrimination and a diminished sense of privacy. Regarding the measurement of revealed privacy preferences, a key aspect to consider is the monetary value that individuals place on their privacy. Research by [2], [16], [20], and [31] highlights that people often underestimate the value of their privacy when responding to hypothetical scenarios. However, in real-world situations, they demonstrate a greater willingness to pay for privacy. This discrepancy highlights the complexity of privacy valuation and the impact of contextual factors on individuals' privacy-related decisions. Additionally, the research shows that privacy preferences are not static; they can be influenced by how information is framed and presented.

From an economics point of view, a comprehensive survey conducted by [3] provides a valuable overview. The authors examined the economic aspects of privacy, investigating how privacy preferences and behaviors impact market outcomes. They discuss the trade-offs individuals make between the benefits of sharing personal information and the potential costs, such as loss of privacy and exposure to risks. This economic perspective underscores the role of information asymmetry and market dynamics in shaping privacy-related decisions. The authors also highlight the challenges in quantifying the value of privacy, given its subjective and context-dependent nature. According to the study by [32], the value of customer information arises from the ability of Internet firms to identify individual consumers and charge them personalized prices. The researchers examine two scenarios: a confidential regime, where the sale of customer information is prohibited, and a disclosure regime, where one firm can compile and sell a customer list to another firm that uses this information for price discrimination. In the era of online platforms and social media, which have transformed how personal information is shared and perceived, [1] examine the complexities of privacy in the digital age. They argue that while individuals have a strong desire for privacy, achieving it is increasingly complex due to the pervasive nature of digital surveillance and data collection. This work highlights the psychological factors that drive privacy concerns, including the tension between the desire for social connection and the need for personal space. It is also essential to reference how preferences respond to changes in actual( [22], [14]) or perceived economic consequences of sharing data ([19], [24]). The incentives of a digital business to collect and protect users' data is advanced by [12]. They characterize how the company's revenue model shapes its optimal data strategy, including the collection and protection of users' data. From a consumer perspective, the balance between consumer privacy and price discrimination in online retail is evaluated (see [17]).

Many studies use an experimental approach to address privacy concerns from an economic and consumer behavior perspective. This branch of literature examines the value of privacy by measuring individuals' willingness to pay for it ([26], [6], [13], [17]). Evidence suggests that the endowment effect, as well as behavioural and cognitive heuristics, play significant roles in shaping privacy valuations (see [18], [29], [30]). The degree of anonymity affects how people behave, as stated by [34] and [28]. More recently, [21] disentangles the intrinsic and instrumental components of the consumer's privacy preference through an experimental approach.

As we've already mentioned, our research aligns with the fascinating dual-privacy theory proposed by [21], which includes two key components: intrinsic and instrumental. The intrinsic component reflects a consumer's natural preference for privacy, stemming from their desire to control their personal information. Meanwhile, the instrumental component delves into the economic aspects of sharing personal data, weighing the potential costs and benefits of disclosing such information to businesses. According to [21], the intrinsic component is highly subjective and varies among consumers and across different categories of data. Consequently, some consumers may prioritize privacy more than others and are more likely to perceive privacy violations in various online advertising practices, even if their data is not misused. The decision to share personal data is subjective, as consumers weigh the costs and benefits of doing so with online advertising companies ([21]). While benefits include tailored advertisements, drawbacks may involve higher prices or excessive ads once preferences are identified. If consumers perceive a reduction in benefits or an increase in costs, they are more likely to feel that their privacy has been compromised.

The remainder of this paper is structured as follows. Section 2 describes the experiment, including our payoff-neutral shopping environment, the experimental design, and the procedures employed. Section 3 presents the behavioral hypotheses guiding our study. Section 4 examines the experimental findings, first from an aggregate viewpoint and then from an individual perspective. Finally, Section 5 summarizes the conclusions drawn from the study.

# 2 The experiment

In this section, we first describe the elements that comprise the shopping environment. They include consumers (preferences), the shopping site (artificial intelligence that governs the algorithm to set prices and products to sell), and the payoffs to consumers from buying the products at various prices. Then, we describe the payoff matrices that were used in our experiment, which categorize various consumer strategies.

## 2.1 A payoff-neutral shopping environment. Theoretical framework

We consider a simplified shopping environment with one consumer and one firm. At each round, the firm selects which of two products to sell ($A$ or $B$) and at which price (low or high), and the consumer decides whether to buy or not.

### 2.1.1 Consumers

Consumers have types based on their dominant preference: A-dominant and B-dominant. At the beginning of the experiment (round 0), consumers are randomly assigned one of the two types with equal probability and the type (dominant preferences) is kept constant over time.

At each round/round $t \geq 1$, consumers are characterized by a temporal preference and a willingness to pay (WTP). The temporal preference of the consumer in each round depends on their dominant preferences. In the case of A-dominant consumers (AAAB), in each round, they will have temporal preference A with a probability of 0.75, meaning that they tend to prefer products or services that satisfy their A-dominant preferences. However, there is a probability of 0.25 that they will have temporal preference B, indicating that they may sometimes prefer products or services that satisfy their B preferences. On the other hand, B-dominant consumers (ABBB) will have the temporal preference B in each round with a probability of 0.75, and there is also a probability of 0.25 that they will have temporal preference A. Therefore, the temporal preference is not fixed but assigned at each round, either $A$ or $B$.

At round t, consumers are not only characterised by their temporal preference (either A or B), but also by their willingness to pay (WTP), which can be either high or low. Whenever the temporal preference of the consumer at round t coincides with their type, then their WTP is high with a probability of 0.75 and low with a probability of 0.25. This suggests that when a consumer's dominant preferences align with their temporal preference, they are more willing to pay a higher price for a product or service that meets their preferences. However, if the participant's type does not coincide with her temporal preferences, then their WTP is low with a probability of 0.75 and high with a probability of 0.25. This indicates that when a consumer's preferences do not align with their temporal preference, they are less willing to pay a higher price for a product or service that does not meet their type.

At each round, consumers have to decide whether to buy the item offered by the online platform. The participant's payoff is given by the difference between her WTP for the product that is offered to her and the price paid.

### 2.1.2 Online Platform

The online platform employs an algorithm based on artificial intelligence (AI) to decide which item to offer (product $A$ or $B$) and at what price (high or low) to each consumer. In this paper, we are not interested in the optimal design of AI but on how humans interact with an AI system. Hence, we fix a simple Bayesian algorithm that assigns a probability $p_A^t$ that the consumer's type is A. At the initial round ($t = 1$), and given that the AI has no prior information about the consumer's true type, the initial AI belief is $\frac{1}{2}$. At subsequent rounds ($t > 1$), the AI updates its belief about the consumer's type $p_A^t$ recursively using Bayesian updating.

It is reasonable to assume that the belief $p_A^t$ determines the product to offer and the selling price in a monotonic way. That is, the larger the belief $p_A^t$ the larger the probability of offering product $A$ and also the larger the probability of setting a high price. (The algorithm seeks to optimise the product and pricing recommendations to maximise the likelihood of the consumer's engagement and satisfaction, by incorporating the consumer's type and the history of their interactions with the platform.)

It is important to highlight that allowing the AI to discover the consumer's true type can have negative consequences for the consumer. Although it increases the chances that the most preferred item is offered by the shopping site, it also increases the chances that it is sold at a high price. In the following section, we describe how we construct a shopping environment (algorithm and consumer's payoff matrices) where a consumer's welfare is not affected by how much private infomation is learnt by the AI. We call this neutral AI.

#### 2.1.3 Fine-tuning the shopping environment for payoff neutrality

- Artificial Intelligence: At round $t$, the AI chooses to sell item $A$ with probability $p^t_A$ and $B$ with probability $1 - p^t_A$, where $p^t_A$ is the probability (belief) the AI has that the consumer's type is $A$. If the item chosen is $A$, the AI sets the price to be high with probability $p^t_A$ and low with probability $1 - p^t_A$, whereas probabilities are reversed otherwise; i.e., if the item chosen is $B$, the AI sets the price to be high with probability $1 - p^t_A$ and low with probability $p^t_A$. This means that the AI is biased towards offering high prices when it believes the consumer is A-dominant and is offering an $A$ product, but may still offer low prices with some probability just in case the consumer is not A-dominant. The mirror reasoning applies if the AI believes the consumer is B-dominant.

- Consumer's payoff matrices: The consumer receives a positive payoff only when they buy an item that coincides with their temporal preference, that is, A-dominant consumers receive positive payoffs with higher probability from buying product $A$, while B-dominant consumers receive positive payoffs with higher probability from buying product $B$. If a consumer buys a product that does not coincide with their temporal preference, they receive a negative payoff. Moreover, the payoff is higher for consumers who pay a lower price for the item. This means that if a consumer buys the product that aligns with their temporal preference, they will receive a higher payoff if they pay a lower price for it. On the other hand, if a consumer buys a product that does not align with their temporal preference, they will receive a negative payoff regardless of the price they pay.

Having described the structure of AI and consumer's payoff matrices, we now analyse the dynamic feedback between consumer's choices over time and the evolution of belief $p^t_A$. This dynamic process will depend on consumer's choices, i.e., on consumer's strategies. We consider a class of strategies that depends on consumer's current payoff.

Specifically, consider the following two strategies: purchase whenever the quoted price is less than or equal to the WTP, and (ii) purchase only when you stand to make the largest possible profit.

It is not easy for consumers to determine the best strategy to use when interacting with the platform, as they might not have access to the AI's internal functioning. The AI's belief about the consumer's type (dominant preferences) is based on the history of their interactions with the platform[2]. As a result, consumers may not be aware of how their past actions and outcomes

[2]This might be advantageous to the algorithm, and the algorithm may exploit it to extract higher profits from

have influenced the algorithm's updating process, making it difficult to predict the algorithm's future recommendations and adapt their behaviour accordingly. The above set of strategies may appear simple, but may actually fit real consumers' behaviour, given the complexity of the decision environment. We can model a strategy as $4 \times 4$ matrix where each element is denoted by $\delta_{ij}^{kl}$, such that $i$ denotes the product $a$ or $b$; $j$ denotes the price of the product, $h$ or $l$; $k$ denotes the temporal preference of the consumer, either $A$ or $B$; and $l$ denotes the consumer's willingness-to-pay (WTP) for the product, $H$ or $L$. If $\delta_{ij}^{kl} = 1$, it means that the consumer buys product $i$ at price $j$ when her type is $k$ and her WTP is $l$. On the other hand, if $\delta_{ij}^{kl} = 0$, the consumer does not buy the product. Next table 1 provides the notation of a strategy:

Table 1: Consumer strategy

| | AH | AL | BH | BL |
|---|---|---|---|---|
| ah | $\delta_{ah}^{AH}$ | $\delta_{ah}^{AL}$ | $\delta_{ah}^{BH}$ | $\delta_{ah}^{BL}$ |
| al | $\delta_{al}^{AH}$ | $\delta_{al}^{AL}$ | $\delta_{al}^{BH}$ | $\delta_{al}^{BL}$ |
| bh | $\delta_{bh}^{AH}$ | $\delta_{bh}^{AL}$ | $\delta_{bh}^{BH}$ | $\delta_{bh}^{BL}$ |
| bl | $\delta_{bl}^{AH}$ | $\delta_{bl}^{AL}$ | $\delta_{bl}^{BH}$ | $\delta_{bl}^{BL}$ |

Consider the strategy "purchase whenever the quoted price is less than or equal to the WTP". This strategy entails that any product that provides a positive payoff will prompt the consumer to acquire it. If the temporal preference coincides with the product, the consumer will purchase it at any price point, regardless of their willingness to pay.

Table 2 corresponds to the strategy under the assumption that consumer gets positive payoff when role and product coincides.

Table 2: Strategy "purchase whenever the quoted price is less than or equal to the WTP"

| | AH | AL | BH | BL |
|---|---|---|---|---|
| ah | 1 | 1 | 0 | 0 |
| al | 1 | 1 | 0 | 0 |
| bh | 0 | 0 | 1 | 1 |
| bl | 0 | 0 | 1 | 1 |

Consider now the strategy "purchase only when you stand to make the largest possible profit". This strategy has the following form (see Table 3):

The above two strategies are simple and only take into account the current payoff. Bear in mind that the AI updates the probability of facing an A-dominant consumer by considering the history of actions made by the consumer. Namely, before $t = 1$, it assumes that types are equally

consumers.

Table 3: Strategy "purchase only when you stand to make the largest possible profit"

| | AH | AL | BH | BL |
|---|---|---|---|---|
| ah | 0 | 0 | 0 | 0 |
| al | 1 | 0 | 0 | 0 |
| bh | 0 | 0 | 0 | 0 |
| bl | 0 | 0 | 1 | 0 |

represented, hence with equal probability individuals have A-dominant and B-dominant types. This means that the initial probability of offering a type $A$ product, denoted by $p_A^0$, is equal to $\frac{1}{2}$, the initial probability the AI attaches to type A for the consumer. At $t \geq 1$, if the consumer purchases items of a certain type, the AI increases the probability of assuming that the consumer is of that type. If the consumer does not purchase items of a certain type, the AI increases the probability of assuming that the consumer is of the opposite type. In other words, the AI tries to learn both when the consumer makes a purchase and when they do not. Formally, the belief the AI has that consumer has A-dominant type at the end of round $t$ denoted by $p_A^t$ is:

$$p_A^t = \frac{C + \sum_{j=1}^{t}[\delta_{a_jc} + \delta_{b_j\bar{c}}]}{2C + \sum_{j=1}^{t}[\delta_{a_jc} + \delta_{a_j\bar{c}} + \delta_{b_j\bar{c}} + \delta_{b_jc}]}$$

where $\delta_{a_jc}$ is equal to 1 if at round $j$ the offered product is $a$ and the consumer purchases it, and equal to zero otherwise; $\delta_{a_j\bar{c}}$ is equal to 1 if at round $j$ the offered product is $a$ and the consumer does not purchase it, and equal to zero otherwise; $\delta_{b_jc}$ is equal to 1 if at round $j$ the offered product is $b$ and the consumer purchases it, and equal to zero otherwise; $\delta_{b_j\bar{c}}$ is equal to 1 at round $j$ the offered product is $b$ and the consumer does not purchase it, and equal to zero otherwise; and $C$ is a positive constant that controls the velocity of the AI learning. In our experiment, we set $C = 1$.

Considering the AI's updating process, we analyse the convergence of the AI's belief for both strategies. It becomes evident that if the consumer purchases the product whenever her payoff is positive, a complete revelation of her true type occurs. However, if the consumer only buys when she stands to gain the maximum profit, the revelation is relatively limited. In this case, the AI's belief remains nearly constant over time, close to 0.5. Figure 1 illustrates the difference between the two strategies. It is clear that when the consumer buys the product whenever her payoff is positive, a full revelation of her true type is generated. However, when the consumer buys only when she stands to make the largest possible profit, the revelation is relative small, as the AI's belief remains almost flat over time close to 0.5.

The last phase involves selecting specific values for high and low willingness-to-pay as well as high and low prices. This selection is done in a manner that ensures the long-term payoff

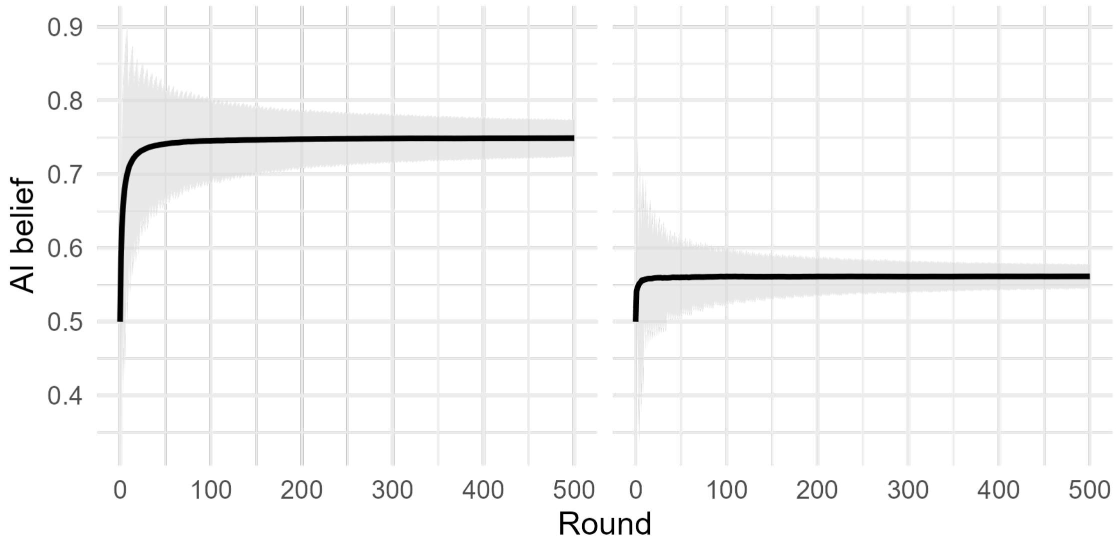


Figure 1: Convergence of AI beliefs as a function of the number of purchases (rounds, x-axis) when shoppers follow the strategies detailed in Table 2 (left panel), purchasing whenever the quoted price is less than or equal to their WTP, and Table 3 (right panel), purchasing only when they stand to make the largest possible profit. The y-axis represents the AI's belief about the buyer's true type. In the left panel, complete revelation of the true type occurs, whereas in the right panel, revelation is limited, stabilizing near 0.5.

for a consumer, employing the strategy of making a purchase whenever the quoted price is less than or equal to their WTP, is equivalent to the payoff obtained from the strategy of purchasing only when the consumer stands to make the maximum possible profit. To achieve this, we need to carefully define and set the values for WTP and prices in a way that aligns the outcomes of the two strategies. Consider the following consumer's payoff matrix (see Table 4):

Table 4: Payoffs for consumer type A-dominant.

| | | Role A | | Role B | |
|---|---|---|---|---|---|
| | | High WTP | Low WTP | High WTP | Low WTP |
| Item a | high price | 0.20 | 0.02 | -0.20 | -2.00 |
| | low price | 2.00 | 0.20 | -0.02 | -0.20 |
| Item b | high price | -0.20 | -2.00 | 0.20 | 0.20 |
| | low price | -0.02 | -0.20 | 2.00 | 0.02 |

A formal analysis in Appendix A shows the steady state of the AI'S belief for any strategy and the corresponding long run payoff. In particular, for the above payoff matrix, our analysis reveals that when the consumer follows a strategy purchase whenever the quoted price is less than or equals to the WTP, the asymptotic AI belief equals 0.75, and the corresponding average payment per round converges to 0.3148437. However, if the consumer adopts the strategy equals to that of the strategy to purchase only when she stands to make the largest possible profit, then the asymptotic AI belief drops to 0.5615528 and the corresponding average payment per round becomes 0.3077641 in the long run. In our case, if consumers follow the former strategy, they reveal their real type, whereas following the latter strategy results in little to no revelation.

This shopping environment is payoff-neutral since both strategies result in the same long-run payoff, but differ in the level of privacy revelation. In the next section, we present the experimental design and procedures.

### 2.2 Experimental design

In the lab, experimental subjects make purchasing decisions in two blocks of 30 rounds each. At the beginning of the experiment, they are informed that there will be two blocks but the details about Block 2 are provided when Block 1 is over. The total payoff from the experiment is the sum of the payoffs over the 60 rounds.

At the beginning of each block, experimental subjects are assigned a type with equal probability (50 per cent chance of being A-dominant or B-dominant), and the type is kept constant over the 30 rounds of the block. The consumer's type is private information for the consumer and they are informed that the shopping platform will never be informed about the private type of the consumer.

In each round, and in order to make purchases, experimental subjects need to accept cookies. When asked to accept/reject them, they are told that the information stored in the cookie will be used by the shopping site to "offer the most appropriate buying opportunities based on their consumer profile."

If a cookie is rejected, the consumer will not have the chance to buy any item and proceeds to the next round. Upon acceptance, the experimental subject is informed about their temporal preference and willingness-to-pay in the round. (They are derived following the procedures outlined in subsection 2.1.3, considering both the AI beliefs and the customer's type.) Once temporal preferences and WTPs are assigned, the experimental subject is informed about the item and price on sale and must decide whether to buy it or not. If the subject decides to buy the item, their payoffs in the round will be realised depending on their assigned role, WTP, the item and the price according to Table 4 if the subject is type $A$ and using the mirror table if she is type $B$. If the subject decides not to buy, their payoff in the stage will be zero.

This process is repeated 30 rounds and the monetary payoffs to experimental subjects from each block is the sum of the payoffs accumulated across the 30 rounds of the block. At the end of Block 1, and before experimental subjects are given the opportunity to have another shopping experience for 30 rounds, Block 2, they receive additional information. It is in the content of this information where treatments differ. There are four treatments in the experiment (see Table 5).. In the Appendix, there is a detail account of the information that experimental subjects received in each of the four treatments between blocks.

| | Obs. | Information provided to subjects between Blocks 1 and 2 |
|---|---|---|
| Control | 50 | The online store AI computes a belief about the consumer's type |
| Awareness | 50 | Control + guess of AI belief + true value of AI's computed belief |
| Info | 50 | Awareness + Info on how the AI computes the belief |
| APP | 50 | Awareness + APP to conceal true type from AI |

Table 5: Experimental treatments

Treatment Control. Subjects are reminded that although their type is private information, e.g. that the online store will never know it, cookie acceptance allows the online store to observe their purchasing behavior. This information is used by artificial intelligence to compute and update a belief about the consumer's type (whether the consumer is predominantly of type $A$ or predominantly of type $B$).

This treatment simply informs experimental subjects that the AI computes a belief about their type, and nothing else. In particular, the decision maker is not informed about how the AI computes beliefs and how these beliefs are used by the shopping site to pick which items to sell and at which price in the next rounds. Once experimental subjects are reminded of what the AI does with the information stored in the cookies, they are invited to have a second shopping experience over another 30 rounds.

Notice that this is the weakest awareness strategy and is our baseline treatment aimed at observing shopping behavior when consumers are aware of cookies but know nothing about how the online platforms use them (this setting mimics real world conditions when shopping online).

Treatment Awareness. Experimental subjects receive the same information as in treatment Control, and in addition they are asked to guess which percentage of their private information has been revealed to the AI through their shopping behavior. This guessing exercise is implemented by asking subjects a number in the interval [0,100], where 100 means full revelation and 0 no revelation at all. This task is incentivized, as experimental subjects could add up to 3 EURO depending on how close the guess was to the actual figure. [3]

Once experimental subjects enter their guess, they learn the actual percentage of their private information revealed to the AI and they are also informed of a feature of the shopping

[3]Given the Bayesian updating process used by the AI, no revelation means that the belief equals its starting value, 0.5, and full revelation means that the AI's belief is $p = 0.75$ of the consumer being their true type. More specifically, a guess $0 \leq x \leq 100$ translates to the $p$-scale using the formula $0.5 + 0.25x/100$.

environment: that the revelation of private information does not harm consumers as it has no impact on individual payoffs. This piece of information is critical to our design, as we are interested in analyzing intrinsic preferences for privacy. Specifically, we tell experimental subjects that "in previous sessions, participants who revealed a lot of private information to the algorithm obtained profits in Block 1 similar to those who revealed little private information."

The aim of this treatment is to explore whether being informed of how much private information is revealed through shopping behavior and what its impact is on their payoffs affects customers' future shopping behavior, i.e., whether failure to do so can stem from different motivations. The first is that experimental subjects do not have any privacy concerns in the lab as their private information is alien to them and exogenously decided by a computer. Besides, it could be that experimental subjects do have privacy concerns even in this artificial environment but do not know how to conceal their private information from the AI. The remaining two treatments are designed to address this issue. The starting point is the conditions as in Treatment Awareness, with an additional feature.

<u>Treatment Info</u>. The additional feature that this treatment adds to the Treatment Awareness is that experimental subjects are informed in plain language about the way in which the AI updates and uses the information stored in the cookies. We were aware that providing subjects with the mathematical formulae was not enough to increase understanding of how the AI learns. So we opted for including a description of which shopping decisions lead the AI to increase the probability of being of a certain type. [4]

Hence, we informed participants that if an experimental subject bought "item $A$" or did not buy "item $B$", then the AI increases the probability that the decision maker is mostly type $A$. Similarly, if the experimental subject bought "item $B$" or did not buy "item $A$," then the algorithm increases the probability that the decision maker was mostly type $B$. Experimental subjects, if interested, are given the possibility of seeing the exact mathematical formula by clicking a button.

Once we explained how the AI beliefs are updated, participants are told, again in plain language, how the AI decides which shopping basket to offer you and at what price.

- If with a high probability the AI considered you to be of the type "mostly A: AAAB", then with a higher probability the shopping firm offered you the "a-basket" at a high price.
- If with a high probability he considered that you were of the "mostly B: BBBA" type, then with a higher probability he offered you the "b basket" at a high price.

<u>Treatment App</u>. In this treatment, the additional feature is that experimental subjects have the opportunity to download for free an APP that advises them whether to purchase the item

[4] This two-step explanation of the AI's functioning in plain language was tested in pilot sessions, which demonstrated its effectiveness in reducing the disclosure of private information in the second block.

based on the consumer type and willingness to pay in the round. The strategy followed by the APP is to safeguard their privacy regarding the consumer's true type. This APP is not an adviser, but a broker, as the experimental subject will necessarily follow the recommendation in all rounds.

In all treatments, participants are informed about their total payoff from Block 1. In treatments Awareness, Info and App, experimental subjects, subjects are also informed about the payoff-neutral feature of the shopping environment. And finally they are given the opportunity to enjoy another shopping experience for additional 30 rounds in the same online store and with the same artificial intelligence algorithm will follow. The consumer's type is reset, as this will be randomly assigned again with a 50-50 probability.

## 2.3 Experimental procedures

The experiment was run at the Laboratory for Research in Experimental Economics (LINEEX) at the University of Valencia in 2022. Participants were recruited from the subject pool of the laboratory using LINEEX recruitment system. We conducted four treatments with 50 participants per treatment, resulting in 200 participants across all four treatments. Of these participants, 59% were women and 41% were men. The average age was 20.8 years, and 60% of the participants were studying Social and Legal Sciences. The instructions -which are provided in the appendix- were read aloud at the beginning of each session. Participants then answered nine self-evaluation questions to assess their understanding of the shopping environment. Each session lasted approximately 50 minutes. The average payment was 12.29 Euros. At the end of the experiment, and while participants were waiting for their payment, they fulfilled a short sociodemographic questionnaire -including age, gender, degree-, a 10-item version of the Big-Five personality test ([27]), Cognitive Reflection Test (CRT), a question to elicit risk aversion and questions about their online shopping habits.

Two features of the design require detailed explanation. First, the number of observations needed per treatment to pick treatment effects and second the number of rounds per treatment.

Regarding the power analysis, we ran some pilot sessions which informed us about the magnitude of the standard deviation of the AI beliefs $\sigma$ as well as the effect size $\Delta$. Specifically, we got $\sigma = 23.30$ and $\Delta = 20.67$. Plugging for power $1 - \beta = 0.8$ and a significance level of $\alpha = 0.05$ in equation $n = \frac{(z_{(\alpha/2)} + z_\beta)^2 \sigma^2}{\Delta^2}$, we get an effect size of 20.67 units with $n = 10$. We took a conservative approach and considered an effect size of 10 units, which results in a required sample size of $n = 50$. This sample size of 50 participants is also a standard convention in behavioral sciences ([9]), looking for a proper balance between statistical power and practical feasibility.

The second important consideration in our design is the number of rounds per block. We based our choice in two pieces of information: simulations and experimental data from pilot sessions. Simulations are illustrated in Figure 2 (left and central panels) and show that for the two previously discussed shopping strategies, the AI-beliefs settle around some limit value after a small number of periods (less than 10). Second, the evolution of the AI beliefs in the pilot sessions also reached a flat plateau in less than 30 rounds. In the right panel in Figure 2, we depict the evolution of the AI-belief aggregating data for Block 1 across all treatments.

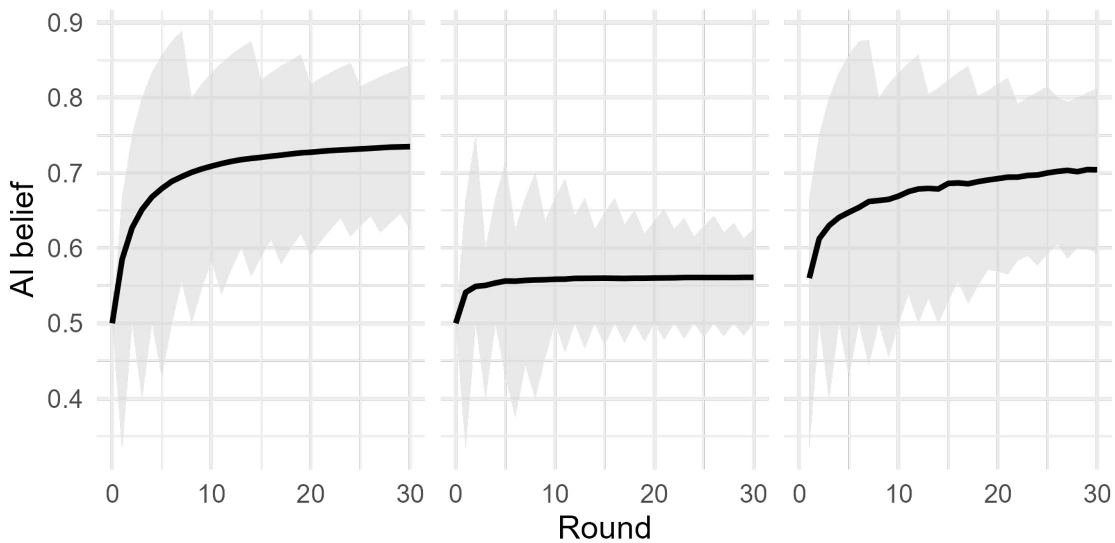


Figure 2: 30 rounds simulation of the evolution of AI beliefs when shoppers follow shopping strategy "Purchase whenever the quoted price is less than or equal to the WTP", (left panel), and strategy "Purchase only when they stand to make the largest possible profit (middle panel), and evolution of AI beliefs in Block 1 aggregating all treatments (right panel)

The final key feature of our procedure involves a systematic approach to comparing the behaviors of our subjects. To effectively measure changes in individual behavior over time, we conduct two distinct blocks, each comprising 30 rounds of interaction. Block 1 is specifically designed to demonstrate individuals' natural inclination to share a considerable amount of private information. In Block 2, participants are prompted to reflect on their experiences and estimate the volume of personal information they believe they have shared. This action aims to enhance their awareness of their tendency to disclose excessively and the implications of their self-disclosure habits. This raises awareness of their behavior and encourages discussion on how to address it.

We introduce different treatments, each designed to implement a unique mechanism aimed at facilitating potential changes in behavior. This multifaceted approach not only enhances our understanding of how different factors may influence behavioral shifts but also provides a comprehensive framework for evaluating individual responses across varying conditions.

The 2-block design allows us to perform a within-subject comparison, enabling us to analyze and contrast the behaviors exhibited by the same participant in Block 1 with those observed in Block 2. But it also complicates the evaluation of the treatment effects as the behavior in the second block might be affected by the experience in Block 1. It is important to note that, after controlling for participants' performance in Block 1 (in terms of both revealed information and earnings), any performance in Block 2 can be fully attributed to the treatment conditions. Linear regression models at the individual level reveal that, as expected, only the treatment variable is statistically significant for revealed information. In contrast, neither the treatment nor performance in Block 1 is statistically significant for earnings (see Appendix B).

# 3 Hypotheses

Prior findings in the literature show that individuals state strong privacy preferences in surveys yet reveal large amounts of private data in practice ([25]). Consistent with this evidence, we hypothesize that in our shopping environment with induced private information, we will observe this privacy paradox. [5]

**Hypothesis 1:** *In our shopping environment, consumers do exhibit behavior consistent with the privacy paradox.*

We split this hypothesis into two, these referring to the two defining features of the privacy paradox.

**Hypothesis 1a:** *In our shopping environment, consumers do reveal large amounts of private information.*

**Hypothesis 1b:** *In our shopping environment, consumers keep revealing large amounts of private information after being briefed about the existence of the AI, which computes how much private information their shopping behavior reveals.*

Once we develop a shopping environment that exhibits the disclosure of large amounts of private information even after participants are briefed in plain language about the existence of the AI, which learns about the private information, we use it to deepen our understanding of the nature preferences for privacy.

The literature on the preferences for privacy considers several explanations for the privacy paradox. Confusion can be ruled out in our experiment as the features and functioning of the shopping environment are carefully explained in the experimental sessions, and experimental

[5] We do not ask participants directly about their privacy concerns, as is commonly done in survey-based studies. Instead, we infer the presence of strong privacy preferences from participants' behavioural responses across treatments. We will show that participants reveal a large amount of private information, they are aware of that, and they reduce the exposure of private information when they are given clues on how the AI works or they have access to an application that limits the transmission of private information to the AI.

subjects passed a comprehension test. A second explanation is wrong beliefs, i.e. that consumers are not aware of the large amount of private information that their shopping behavior reveals. The literature also shows that there might be two distinct components in the preferences for privacy: the instrumental one -related to the potential monetary consequences of the disclosure of private information- and the intrinsic component -related to the value of privacy per se ([21]).

We hypothesize that the intrinsic component of privacy is immune to wrong beliefs. This is the content of Hypothesis 2, which takes advantage of the payoff-neutrality feature of our shopping environment.

**Hypothesis 2:** *Participants keep revealing large amounts of private information even after they are made fully aware of the large amount of private information they are revealing, even in environments where consumers are aware that there are no monetary gains from trading private information.*

A skeptical reader might argue that the large amount of private information revealed by participants in our experiment says nothing about their preferences for privacy as the nature of the "private information" is artificial and induced in the lab. In addition, other readers might argue that in a shopping environment where disclosing private information does not harm monetary payoffs, it is weakly dominant to disclose any amount of private information, and therefore the level of disclosure cannot be informative about the preferences for privacy.

To answer these two claims, rather than asking subjects about their preferences for privacy ([3]), we hypothesize on the impact of the introduction of two mechanisms on the revelation of private information, aimed at combating the third reason that might be behind the privacy paradox: resigned privacy ([10], [11]). If there were no intrinsic privacy concerns at all in the experiment, then no mechanism aimed at helping participants reduce the sharing of private info would have an impact on the amount of private information revealed.

**Hypothesis 3:** *Intrinsic preferences for private concerns do exist.*

We divide this hypothesis into two:

**Hypothesis 3a:** *In our shopping environment, the amount of private information revealed decreases when consumers are briefed about how the AI works (explainable AI).*

**Hypothesis 3b:** *In our shopping environment, the amount of private information revealed decreases when consumers are offered a privacy APP.*

# 4 Experimental results

In this section, we analyze in first place the experimental results using aggregate data. Then, we perform an analysis at the individual level.

## 4.1 Aggregate analysis

Table 6 provides information on aggregate behavior by treatment and blocks of 30 rounds. The first two columns display the average payoffs per block, averaged across subjects, (Block 1 and Block 2 respectively), and the third one a parametric paired t-test of differences across blocks. The next three columns give the percentage of private information revealed (the AI beliefs in a scale 0-100)[6] at the end of each block, averaged across subjects (Block 1 and 2) and the p-values of the paired t-test of differences across blocks. The last column contains the average value across subjects of the guess made by each subject and how much the AI belief has learned about their type.

Table 6: Statistical summary of payoffs, AI beliefs and guesses by treatment.

| | Payoffs | | | AI beliefs | | | Individual |
|---|---|---|---|---|---|---|---|
| Treatment | Block 1 | Block 2 | p-values | Block 1 | Block 2 | p-values | guesses |
| Control | 8.73 | 9.31 | 0.4341 | 73.5 | 75.0 | 0.7547 | - |
| | (3.68) | (3.24) | | (24.8) | (25.6) | | ( - ) |
| Awareness | 9.02 | 9.75 | 0.3350 | 74.4 | 73.8 | 0.8908 | 77.4 |
| | (4.23) | (3.60) | | (27.3) | (28.2) | | (19.5) |
| Info | 8.74 | 9.05 | 0.6897 | 79.6 | 64.1 | 0.0072 | 73.7 |
| | (3.77) | (3.79) | | (27.0) | (32.2) | | (21.2) |
| App | 9.03 | 9.77 | 0.3464 | 78.7 | 57.2 | $< 0.0001$ | 75.5 |
| | (3.79) | (3.99) | | (23.2) | (34.0) | | (20.1) |

Source: compiled by the authors. Averages and standard deviations (in brackets) have been computed by block and treatment. P-values have been computed using t-test for paired samples.

We first focus on the amount of private information revealed by subjects through their shopping behavior (Hypothesis 1a) at the end of Block 1. The fourth column of Table 6 contains the percentage of private information learnt by the AI at the end of round 30 for each treatment. We see that percentages are in the interval (70 - 80 percent) and that there are no significant differences between them[7]. When aggregated, the percentage of private information revealed to the AI is around 75 per cent (76.6), significantly different from zero (p-value $< 0.0001$) and closer to one hundred than to zero (p-value $< 0.0001$; see left panel of Figure 3). Hence, our first result is as follows:

[6] Where 100 means full knowledge of true type and 0 no knowledge at all.

[7] The p-value of a paired t-test of equal means is 0.7547.

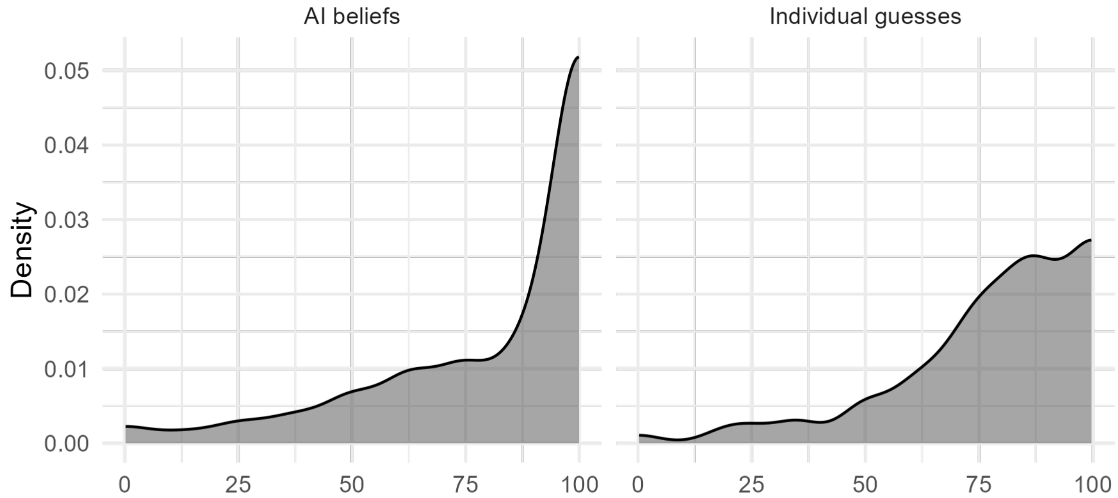


Figure 3: Estimated distributions of AI learning (left panel) and subjects' guesses (right panel) at the end of Block 1, with all observations pooled except for the Control treatment. The x-axis represents the percentage of the buyer's true type learned by the AI (left panel) and the subjects' individual guesses about this percentage (right panel). The y-axis indicates the density of individuals in each value.

**Result 1:** In our online shop, consumers do reveal large amounts of private information

To claim that behavior is consistent with the privacy paradox, we need to show that subjects keep showing the same levels of private information revelation after being briefed about the existence of an AI that computes how much private information the shopping behavior reveals (Hypothesis 1b). To answer this, we focus on the Control treatment and show that there are no significant differences between the percentage of information revealed at the end of Block 1 (round 30), 73.5 percent and that at the end of Block 2 (round 60), 75.0 percent. This is the case as the p-value of equal means is 0.7547. Hence, our next result is as follows:

**Result 2:** In our online shopping environment, consumers keep revealing the same amount of private information after being informed of the existence of an AI that computes how much private information their behaviour reveals.

We now test the payoff-neutrality of the shopping environment. Focus first in the Control treatment over 30 rounds (at the end of Block 1). A regression analysis of the accumulated payoff on the amount of revelation yields a slope coefficient of 0.03, with a p-value = 0.1659. This features allowed us to say to participants in treatments Awareness, Info and App that "in previous sessions, subjects who had revealed low levels of private info obtained monetary payoffs similar to those received by subjects who revealed large levels of private info".

In fact, we now show that the shopping environment's payoff neutrality holds for all observations of our experimental sessions. The regression analysis shows that the slope is not statistically significant (p-value: 0.1936).

**Result 3:** The experimental implementation of our shopping site shows that monetary payoffs are independent of the amount of private information revealed to the AI.

Recall that in treatment Awareness, Info, and App, at the end of Block 1, subjects were asked to guess (through an incentivized task) how much private information they have revealed (see right panel of Figure 3). It is only after they have completed the guess task that they were informed about their true level of private information disclosed and that results from previous sessions showed that the level of private information revealed had no impact on participants' payoffs.

Table 6 above includes participants' guesses. We find no significant differences between guesses and the amount of private information actually revealed to the AI in any of the treatments at the end of Block 1. The p-values corresponding to paired t-tests of equal means are 0.4594, 0.1200, and 0.3325 for, respectively, the Awareness, Info, and App treatments. Figure 4 illustrates the histogram of the AI learning for Block 1 across all treatments except for Control, as well as the individual guesses. We see that both distributions are similar.

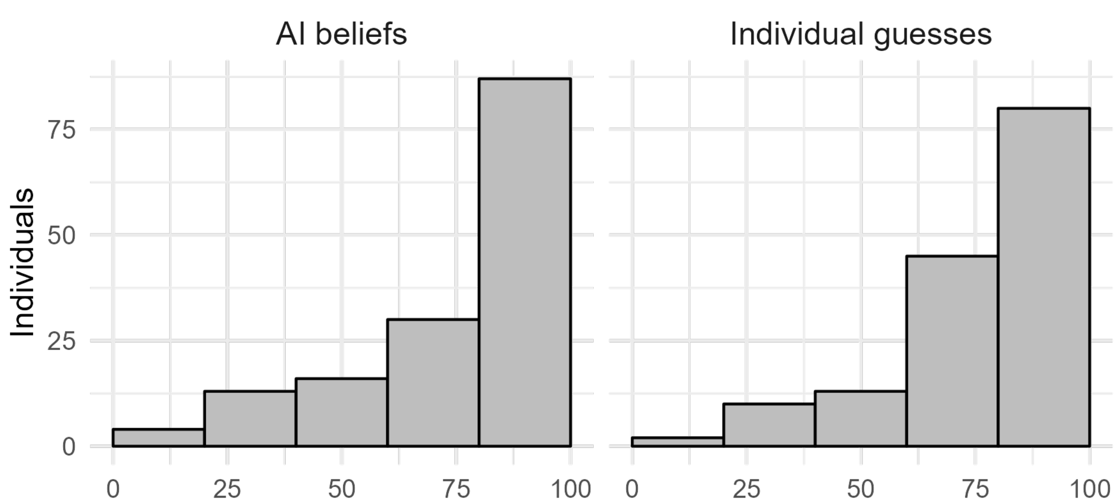


Figure 4: Histograms of AI beliefs of Block 1 of all treatments but not Control and the individual guesses.

This gives rise to our next result:

**Result 4:** Participants know that they were revealing large amounts of private information to the AI.

Focusing on Treatment Awareness, Table 6 shows that the levels of private information revealed across blocks are not significantly different. Furthermore, Figure 5 provides a detailed representation of the AI beliefs histogram for both blocks within the awareness treatment. Notably, the patterns observed in the histograms of the two blocks are similar, suggesting a consistency in the participants' behavior. This observation leads us to conclude our next finding, which substantiates Hypothesis 2.

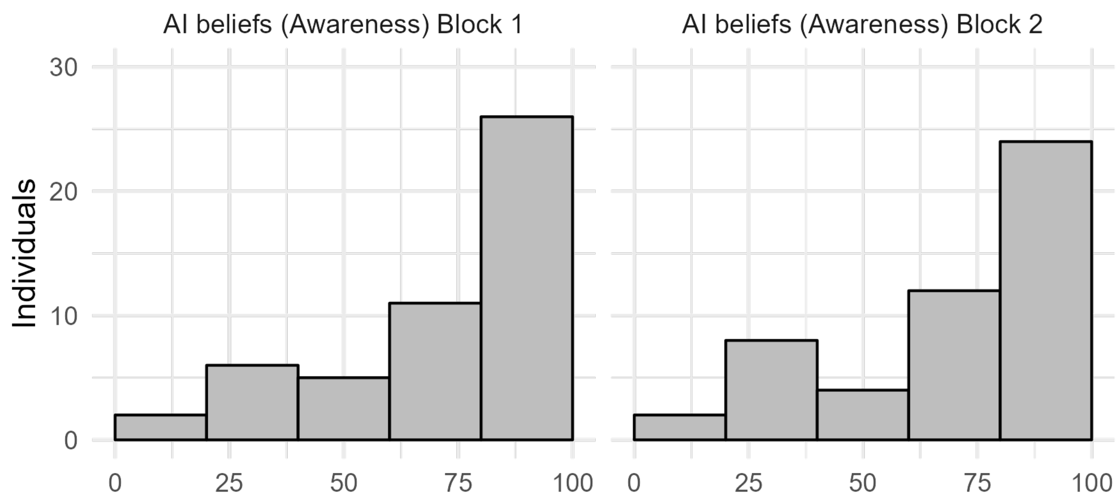


Figure 5: Histograms of AI beliefs of Awareness treatment in Block 1 and Block 2.

**Result 5:** In shopping environments where trading private information offers no monetary benefit, buyers continue to disclose private information even after realizing how much their shopping behavior reveals.

Lastly, we test Hypothesis 3. The last two treatments are designed to check whether the levels of private information revelation across blocks change as subjects are told how the AI works (Treatment Info) and as they are offered a privacy APP (Treatment App).

Table 6 above shows that in Treatment Info, there are significant differences between the private information revealed across blocks, i.e., at the end of Blocks 1 and 2. The percentage of private information decreases from 79.6 to 64.1. This difference is significant at the 1 percent level. This supports our next two results:

**Result 6:** When subjects are told how the AI works, they successfully manage to adapt their shopping behavior to lower the levels of private information revealed to the AI even though they are aware that it has no effect on their monetary payoffs.

We can offer a more comprehensive analysis of how the AI's beliefs evolve between Block 1 and Block 2, specifically in relation to the two treatments: Info and App. To illustrate these changes, we present histograms in Figures 6 and 7, each displaying the distribution of beliefs under the respective treatments. These visual representations allow us to better understand the behavior of individuals in each scenario.

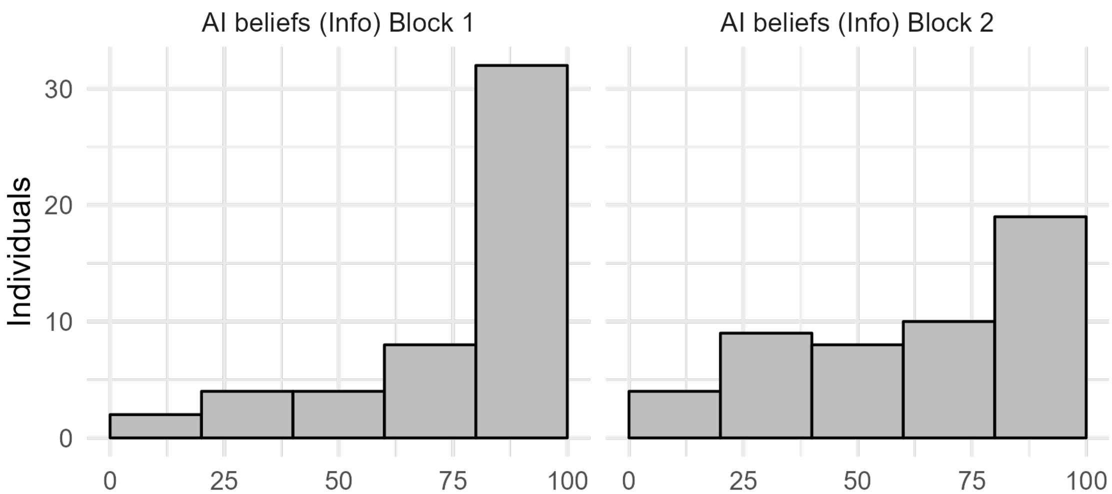


Figure 6: Histograms of AI beliefs of Info treatment in Block 1 and Block 2.

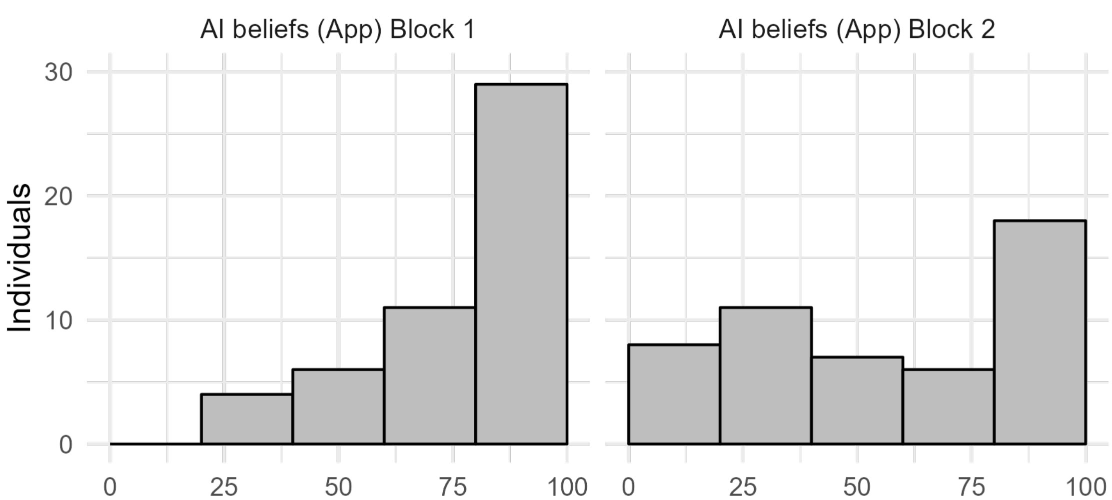


Figure 7: Histograms of AI beliefs of App treatment in Block 1 and Block 2.

Finally, Table 6 above shows that when experimental subjects are offered a privacy APP, the levels of private information revelation decrease from 78.7 to 57.2. This difference is significant at a one percent level (p-value $< 0.0001$). This is achieved because some participants accept the use of the APP (one third of participants, a figure significantly larger than 10 percent, p-value $< 0.0001$)[8]. The level of private information revealed by those who choose to use the APP automatically lowers to 20.21 percent. However, those who did not accept the use of the APP maintained the same level of private information revelation in Block 2 (74.63 percent) as in Block 1 (78.87 percent, with the paired t-test of equal means yielding a p-value of 0.5236).

**Result 7:** When subjects are offered a Privacy APP, one third of subjects commit to using it. Those who do not use it, keep revealing the same private information.

Figure 8 and Figure 9 present the histogram of the AI beliefs comparing individuals who do not use the APP (on the left) to those who do use it (on the right) where the x-axis represents the learning level of the AI first in Block 1 and then in Block 2, respectively. Each histogram provides a visual representation of the varying AI beliefs held by these two distinct groups. It is evident that individuals using the APP exhibit a lower revealing level, with the APP achieving less than 60% of learning in Block 2. This indicates a shift of learning levels from the right to the left for those using the APP. Moreover, there is a compelling observation regarding those individuals who shared their type during Block 1 and subsequently began using the APP. This subset shows a distinct difference in their revelation levels during Block 2, indicating that prior behavior without the APP may influence how they interact with the AI learning process after adopting it. This shift further highlights the importance of understanding the impact of technology on learning beliefs and behaviors.

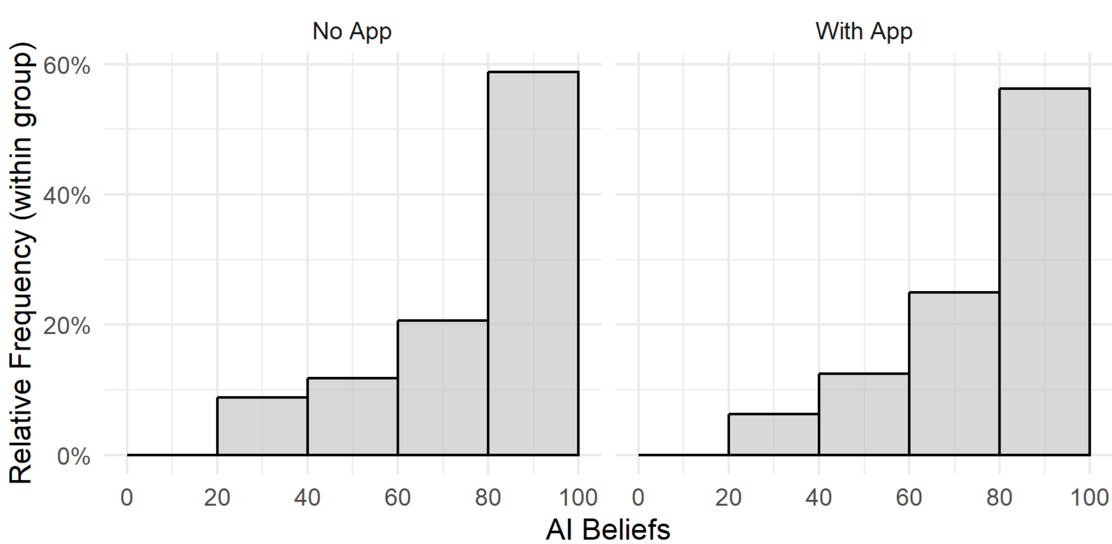


Figure 8: Histograms of AI beliefs in Block 1 for those individuals that do not use the APP (left side) and those that use the APP (right side).

[8]This is a conservative test, because rare events are usually assumed to be selected by less than 5 percent of people.

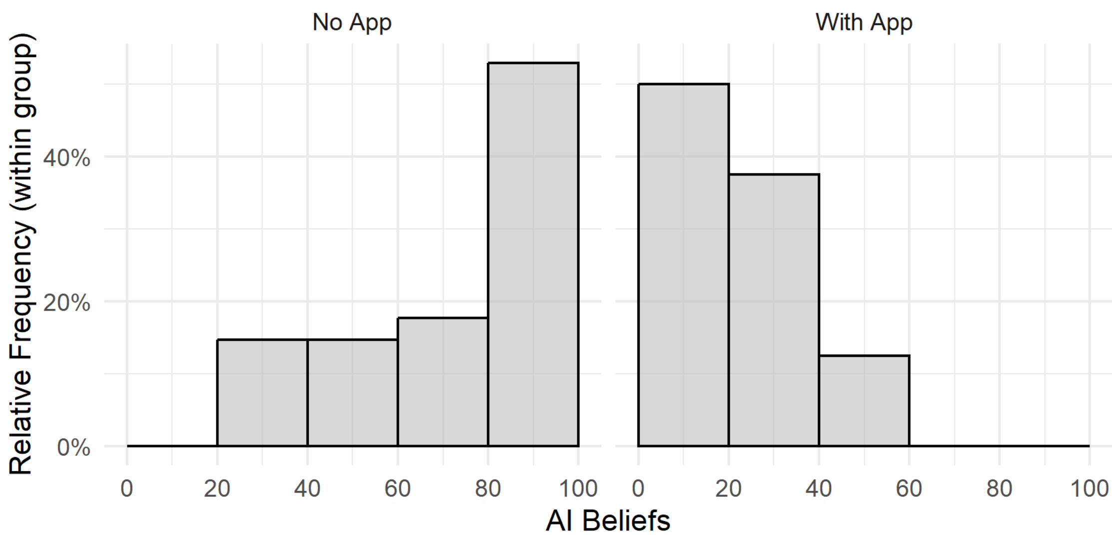


Figure 9: Histograms of AI beliefs in Block 2 for those individuals that do not use the APP (left side) and those that use the APP (right side).

## 4.2 Individual analysis

In this section we perform the analysis of the experimental data at the individual level. We first look at the relationship between information disclosure and monetary payoffs. Figure 10 shows this relationship. For each participant, we have two points, one per block. We highlight the observations in Block 2 in treatments Info or App using triangular-shaped red points.

Two interesting patterns emerge. On the one hand, for any level of private information revealed, the associated payoffs vary at great length. This is reminiscent of the random choices of items and prices by the AI. And on the other hand, we observe a relatively higher density of red triangles in the left side of the figure -i.e. for low levels of privacy disclosure-: 36.5% versus 20.6%, p-value = 0.0003. This shows, using observations at the individual level, that the two mechanisms offered to participants help them reduce their level of private information revealed in Block 2 compared to the information disclosed in the absence of these mechanisms. We next explore whether there are some determinants at the individual level that explain the change in behavior in the presence of these two mechanisms. We use the data from the questionnaire that participants fulfilled while waiting to receive their monetary payoffs.

For Treatment Info, Table 4.2 contains the OLS regression analysis where the dependent variable is the change in behavior, defined as the difference between the amount of information disclosed by a subject in Block 1 minus the amount of information disclosed in Block 2. The independent variables are guess (the guess about the private information disclosed in Block 1), a measure of risk aversion (using the investment problem), gender (whether the subject is female), a measure of cognitive ability (measured by the score in CRT), online (measured by how

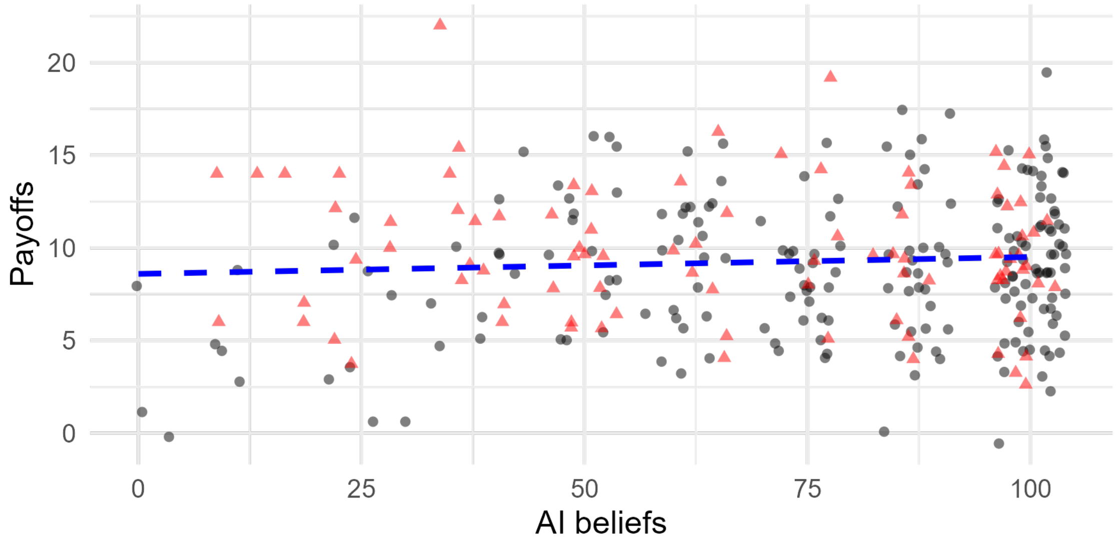


Figure 10: Relationships between payoffs and private information disclosure at the subject level. There are two points for each subject (one point per block). Triangular-shaped, red points correspond to subjects in Block 2 for the Info and App treatments. A regression line has also been added to the plot. Its slope is not statistically significant (p-value: 0.1936).

often a subject engaged in online shopping), and the five variables from BIG-5 (extraversion, agreeableness, conscientiousness, neuroticism, and openness).

| **Variable** | **Estimate** | **Std. Error** | **t value** | **p-value** |
|---|---|---|---|---|
| (Intercept) | 49.97507 | 43.56935 | 0.940 | 0.3528 |
| Guess | 49.38965 | 26.99858 | 1.829 | 0.0750 + |
| Risk | 20.75300 | 49.53890 | 0.419 | 0.6776 |
| Female | -29.10612 | 11.21516 | -2.595 | 0.0133 * |
| CRT | -5.46633 | 5.27096 | -1.037 | 0.3061 |
| Online | -5.86442 | 13.02593 | -0.450 | 0.6551 |
| Extraversion | 0.09138 | 5.15629 | 0.018 | 0.9860 |
| Agreeableness | -1.58893 | 5.46598 | -0.291 | 0.7728 |
| Conscientiousness | -7.65313 | 5.34136 | -1.433 | 0.1599 |
| Neuroticism | 3.98526 | 4.79748 | 0.831 | 0.4112 |
| Openness | -3.77490 | 4.96932 | -0.760 | 0.4520 |
| *Signif. codes:* 0 '***' 0.001 '**' 0.01 '*' 0.05 '+' 0.1 ' ' 1 | | | | |
| Multiple R-squared: 0.3164, Adjusted R-squared: 0.1412 | | | | |

Table 7: Determinants of change in behavior in Treatment Info. OLS

Next table 8 provides the OLS regression for the private App treatment with the same dependent variable and the independent variables as the previous regression, but it also includes a dummy variable that takes the value one if the participant used the APP.

| Variable | Estimate | Std. Error | t value | p-value |
|---|---|---|---|---|
| (Intercept) | -67.942549 | 50.804727 | -1.337 | 0.1891 |
| Guess | 57.885106 | 25.801845 | 2.243 | 0.0308 * |
| Risk | 64.806097 | 45.653144 | 1.420 | 0.1639 |
| Female | -0.008080 | 10.412280 | -0.001 | 0.9994 |
| CRT | -5.176213 | 4.652716 | -1.113 | 0.2729 |
| Online | 6.959833 | 10.205197 | 0.682 | 0.4994 |
| Extraversion | -0.988885 | 5.446531 | -0.182 | 0.8569 |
| Agreeableness | 2.981246 | 4.984313 | 0.598 | 0.5533 |
| Conscientiousness | 0.101465 | 6.194447 | 0.016 | 0.9870 |
| Neuroticism | -0.009286 | 4.996553 | -0.002 | 0.9985 |
| Openness | 1.929414 | 4.956765 | 0.389 | 0.6993 |
| APP | 54.491747 | 10.924642 | 4.988 | $<$ 0.0001 *** |
| *Signif. codes:* 0 '***' 0.001 '**' 0.01 '*' 0.05 '+' 0.1 ' ' 1 | | | | |
| Multiple R-squared: 0.5789, Adjusted R-squared: 0.457 | | | | |

Table 8: Determinants of change in behavior in Treatment App. OLS

We observe that in both mechanisms, the variable "Guess" is positive and significant at the 5 and 10 percent level in 4.2 and 8 respectively. This indicates that participants who believed they had disclosed more private information in Block 1 adjusted their behavior more significantly between the two blocks. This trend also holds true in the App treatment, even after accounting for whether the participant used the privacy APP. In fact, the APP variable is positive and significant at the 1 percent level because its design aims to conceal the participant's private information. The only significant sociodemographic variable identified in the Treatment Information is gender. Analysis of our data indicates that female participants tend to exhibit less willingness to alter their behavior compared to their male counterparts. This pattern suggests that women may be more consistent in their responses, whether they disclose a great deal of information or remain more reserved.

We will now analyze the factors that influence the adoption of the privacy APP in the Treatment App. Table 9 presents the econometric results from a logit model, where the dependent variable is a dummy variable called "APP." This variable takes the value of 1 if the participant adopted the privacy APP. The independent variables are the same as those in the previous econometric analysis, with the addition of the payoff from Block 1 and the AI belief in the first block.

The understanding of who is likely to adopt the APP is quite interesting. Firstly, the benefits received in Block 1 positively influence the likelihood of obtaining the free private APP.

| Variable | Estimate | Std. Error | z value | p-value |
|---|---|---|---|---|
| (Intercept) | 7.66351 | 5.78194 | 1.325 | 0.18503 |
| Guess | 5.01840 | 4.68249 | 1.072 | 0.28384 |
| Risk | -19.60886 | 8.70878 | -2.252 | 0.02435 * |
| Female | 2.24734 | 1.37436 | 1.635 | 0.10201 |
| CRT | -0.74498 | 0.56814 | -1.311 | 0.18977 |
| Online | -1.53234 | 1.49826 | -1.023 | 0.30643 |
| Extraversion | -0.96287 | 0.88622 | -1.086 | 0.27726 |
| Agreeableness | -1.36521 | 0.71495 | -1.910 | 0.05620 + |
| Conscientiousness | -2.65642 | 1.01087 | -2.628 | 0.00859 ** |
| Neuroticism | 1.32719 | 0.70634 | 1.879 | 0.06025 + |
| Openness | 0.41975 | 0.49926 | 0.841 | 0.40049 |
| Payoffs Block 1 | 0.56744 | 0.25243 | 2.248 | 0.02458 * |
| AI beliefs Block 1 | -0.01310 | 0.02917 | -0.449 | 0.65335 |
| *Signif. codes:* 0 '***' 0.001 '**' 0.01 '*' 0.05 '+' 0.1 ' ' 1 | | | | |
| AIC: 58.449 | | | | |

Table 9: Determinants of adoption of the privacy APP. Logit

There are two variables from the Big Five personality traits—agreeableness and conscientiousness—that have an inverse effect on adoption. One could reason that individuals who are more self-aware might not feel the need for additional assistance that the APP provides. However, if stress is a factor that you struggle to manage, then adopting the APP could be beneficial. This explains the positive correlation with the neuroticism variable. Ultimately, risk aversion has a notable influence; individuals who are more cautious about taking risks tend to be less inclined to adopt the APP. One interpretation of this behavior is that for those who are averse to risk, depending on a machine to recommend their actions—especially when they perceive a loss of control—can be unappealing.

Overall, combining data from both interventions, our analysis identifies certain traits of 'privacy lovers'—individuals who actively protect their private information when given the means to do so. We find that these privacy lovers tend to be those who perceive themselves as disclosing large amounts of private information, as well as female and risk-averse participants.

# 5 Conclusions

**Policy implications.** Our experimental results indicate a significant inclination among humans to disclose substantial amounts of personal information in online settings. Moreover, we demonstrate that simply informing individuals about this tendency is insufficient to prevent the inadvertent exposure of private information. These findings underscore the need for innovative strategies to encourage consumers to adopt privacy-preserving online shopping habits[9].

In our study, we examine two strategies (nudges) in a laboratory setting and find that they effectively reduce privacy exposure. Specifically, we find that providing consumers with insight into the inner workings of AI and offering them an APP that automatically safeguards their private information has a positive impact. However, implementing these strategies in real-world scenarios presents a significant challenge.

The first strategy encounters the challenge that not all AI algorithms are explainable. In cases where the inner workings of the AI are not well understood, providing insights into its functionality becomes difficult. Governments have already recognised the importance of explainable artificial intelligence as a means of building trust and confidence in technology. Therefore, we encourage the public sector to continue promoting the development of explainable AI.

The second strategy confronts the challenge that a privacy-focused APP must be tailored to the specific AI algorithm it is intended to work with. This necessitates the development of a unique APP for each AI, resulting in a time-consuming and resource-intensive process. One potential solution to this issue is to advocate for government-led efforts to standardise the design of AI algorithms, allowing for the creation of a universal APP that can be utilised across all AI algorithms.

One key takeaway from our study is that consumers may demand the adoption of privacy-preserving solutions, even without government mandates. An example of this is the competition among search engines like DuckDuckGo and Brave, which distinguish themselves by attracting users with ad-free and tracker-free experiences, making privacy their unique selling point. Our study suggests that integrating an API specifically designed to manage privacy can create a significant advantage for businesses. By enhancing consumer trust and addressing privacy concerns, this approach helps consumers make informed choices when they prioritize their personal data security.

We have identified two nudges, explainable AI or free Privacy APP, that make a difference from a privacy perspective. For future research, it would be beneficial to explore the concepts of "opt-in" and "opt-out" nudges. These approaches influence user consent mechanisms and can play a crucial role in how individuals choose to share their data.

**Strengths and limitations.** A limitation of our research, as with all experimental studies, is the question of external validity. We acknowledge that there is a trade-off between controlling the environment to establish causal relationships and generalising findings to real-world

[9] The General Data Protection Regulation (GDPR) of the European Union emphasizes the importance of privacy by design and by default in Article 25. This article states that organizations must "implement appropriate technical and organizational measures to ensure that, by default, only personal data necessary for each specific purpose of processing is collected" (European Union 2016).

scenarios. In our case, it was imperative to create a shopping environment where the disclosure of private information did not impact monetary rewards, necessitating an experimental design. Furthermore, we set out to observe behaviour under various circumstances, including information and an APP, which again called for an experimental approach.

However, there is another important aspect to consider in our study, which functions as an a fortiori argument. We employed induced preferences, where the nature of the private information was artificially and randomly determined by a computer. This raises questions about whether experimental subjects would develop a preference for preserving the privacy of such extraneous information. Our results demonstrate that, in fact, our experimental subjects did develop a preference for privacy, as we were able to influence their shopping behaviour towards greater privacy preservation through the two nudges. This a fortiori argument suggests that with personal and genuine private information, subjects would likely exhibit an even stronger response to the nudges.

A natural progression from the insights gained in a controlled laboratory study is to investigate how well the payoff-neutral results translate to real markets, where factors such as personalization may play a crucial role. Conducting a field experiment on an actual e-commerce platform or utilizing authentic personal data could represent the next step in our research agenda.

While our findings suggest that lack of access to privacy-preserving tools can significantly influence disclosure behavior, we do not claim that it is the sole or dominant factor. Our design focuses on isolating this specific mechanism, and future research should explore additional contextual and psychological determinants of privacy-related decisions, especially across different digital environments such as social media platforms, where disclosure norms and incentives may differ markedly from those observed in online shopping contexts.

# References


[1] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Secrets and likes: The drive for privacy and the difficulty of achieving it in the digital age. Journal of Consumer Psychology, 30(4):736–758, 2020.

[2] Alessandro Acquisti, Leslie K John, and George Loewenstein. What is privacy worth? The Journal of Legal Studies, 42(2):249–274, 2013.

[3] Alessandro Acquisti, Curtis Taylor, and Liad Wagman. The Economics of Privacy. Journal of Economic Literature, 54(2):442–492, June 2016.

[4] Alessandro Acquisti and Hal R Varian. Conditioning prices on purchase history. Marketing Science, 24(3):367–381, 2005.

[5] Irwin Altman. The environment and social behavior: privacy, personal space, territory, and crowding. 1975.

[6] Volker Benndorf and Hans-Theo Normann. The willingness to sell personal data. The Scandinavian Journal of Economics, 120(4):1260–1278, 2018.

[7] Bettina Berendt, Oliver Günther, and Sarah Spiekermann. Privacy in e-commerce: Stated preferences vs. actual behavior. Communications of the ACM, 48(4):101–106, 2005.

[8] Barry Brown. Studying the internet experience. HP laboratories technical report HPL, 49, 2001.

[9] Jacob Cohen. Statistical power analysis for the behavioral sciences. routledge, 2013.

[10] Nora A Draper. From privacy pragmatist to privacy resigned: Challenging narratives of rational choice in digital privacy debates. Policy & Internet, 9(2):232–251, 2017.

[11] Nora A Draper, Christian Pieter Hoffmann, Christoph Lutz, Giulia Ranzini, and Joseph Turow. Privacy resignation, apathy, and cynicism: Introduction to a special theme, 2024.

[12] Itay P Fainmesser, Andrea Galeotti, and Ruslan Momot. Digital privacy. Management Science, 69(6):3157–3173, 2023.

[13] David Fehrenbach and Carolina Herrando. The effect of customer-perceived value when paying for a product with personal data: A real-life experimental study. Journal of business research, 137:222–232, 2021.

[14] Eleonora Freddi and Ole Christian Wasenden. Privacy during pandemics: Attitudes to public use of personal data. Journal of Behavioral and Experimental Economics, 113:102304, 2024.

[15] Garrett Glasgow, Sarah Butler, and Samantha Iyengar. Survey response bias and the 'privacy paradox': evidence from a discrete choice experiment. Applied Economics Letters, 28(8):625–629, 2021.

[16] Avi Goldfarb and Catherine Tucker. Shifts in privacy concerns. American Economic Review, 102(3):349–353, 2012.

[17] Sinem Hidir and Nikhil Vellodi. Privacy, personalization, and price discrimination. Journal of the European Economic Association, 19(2):1342–1363, 2021.

[18] Britta Hoyer and Dirk van Straaten. Anonymity and self-expression in online rating systems—an experimental analysis. Journal of Behavioral and Experimental Economics, 98:101869, 2022.

[19] Leslie K John, Alessandro Acquisti, and George Loewenstein. Strangers on a plane: Context-dependent willingness to divulge sensitive information. Journal of consumer research, 37(5):858–873, 2011.

[20] Michael Kummer and Patrick Schulte. When private information settles the bill: Money and privacy in google's market for smartphone applications. Management Science, 65(8):3470–3494, 2019.

[21] Tesary Lin. Valuing intrinsic and instrumental preferences for privacy. Marketing Science, 41(4):663–681, 2022.

[22] Kirsten Martin and Helen Nissenbaum. Measuring privacy: An empirical test using context to expose confounding variables. Colum. Sci. & Tech. L. Rev., 18:176, 2016.

[23] Faverio M. Anderson M. McClain, C. and E. Park. How americans view data privacy, pew research center, 2023.

[24] Amalia R Miller and Catherine Tucker. Privacy protection, personalized medicine, and genetic testing. Management Science, 64(10):4648–4668, 2018.

[25] Patricia A Norberg, Daniel R Horne, and David A Horne. The privacy paradox: Personal information disclosure intentions versus behaviors. Journal of consumer affairs, 41(1):100–126, 2007.

[26] Sören Preibusch, Dorothea Kübler, and Alastair R Beresford. Price versus privacy: an experiment into the competitive advantage of collecting less personal information. Electronic Commerce Research, 13:423–455, 2013.

[27] Beatrice Rammstedt and Oliver P John. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. Journal of research in Personality, 41(1):203–212, 2007.

[28] Simeon Schudy and Verena Utikal. 'you must not know about me'—on the willingness to share personal data. Journal of Economic Behavior & Organization, 141:1–13, 2017.

[29] Anya Skatova, Rebecca McDonald, Sinong Ma, and Carsten Maple. Unpacking privacy: Valuation of personal data protection. Plos one, 18(5):e0284581, 2023.

[30] Tianshu Sun, Zhe Yuan, Chunxiao Li, Kaifu Zhang, and Jun Xu. The value of personal data in internet commerce: A high-stakes field experiment on data regulation policy. Management Science, 70(4):2645–2660, 2024.

[31] Huan Tang. The value of privacy: Evidence from online borrowers. Available at SSRN 3880119, 2019.

[32] Curtis R Taylor. Consumer privacy and the market for customer information. RAND Journal of Economics, pages 631–650, 2004.

[33] TRUSTe. Consumer confidence privacy report consumer opinion and business impact, 2014.

[34] Tobias Vorlaufer. Effects of double-anonymity on pro-and anti-social behavior: Experimental evidence from a lab in the field. Journal of Behavioral and Experimental Economics, 81:216–225, 2019.

[35] Samuel Warren and Louis Brandeis. The right to privacy. In Killing the Messenger: 100 Years of Media Criticism, pages 1–21. Columbia University Press, 1989.

[36] Alan F Westin. Privacy and freedom. Washington and Lee Law Review, 25(1):166, 1968.

# 6 Appendix

## 6.1 Appendix A. Steady State

In this section we study the steady state of the dynamic learning process of AI considering a strategy that purchases whenever the quoted price is less than or equal to the WTP, and a strategy that purchases only when it stands to make the largest possible profit. Note that AI updates the probability of facing an A-dominant consumer by considering the history of actions made by the consumers. At $t = 0$, it assumes that profiles are equally represented, hence with equal probability individuals have A-dominant type. This means that the initial probability of offering a type $a$ product, denoted by $p_a^0$, is equal to $\frac{1}{2} = p_A^0$.

The formula for updating the probability of offering product $a$ (assuming that the consumer is A-dominant) after $t-1$ decisions is equivalent to the belief of AI that the consumer has A-dominant type at the end of round $t$ denoted by $p_A^t$ is::

$$p_A^t = \frac{C + \sum_{j=0}^{t-1}[\delta_{a_j c} + \delta_{b_j \bar{c}}]}{2C + \sum_{j=0}^{t-1}[\delta_{a_j c} + \delta_{a_j \bar{c}} + \delta_{b_j \bar{c}} + \delta_{b_j c}]}$$

where $\delta_{a_j c}$ is equal to 1 if at time $j$ the product is $a$ and the consumer purchases it, and equal to zero otherwise; $\delta_{a_j \bar{c}}$ is equal to 1 at time $j$ the product is $a$ and the consumer does not purchase it, and equal to zero otherwise; $\delta_{b_j c}$ is equal to 1 if at time $j$ the product is $b$ and the consumer purchases it, and equal to zero otherwise; $\delta_{b_j \bar{c}}$ is equal to 1 at time $j$ the product is $b$ and the consumer does not purchase it, and equal to zero otherwise. Therefore, the probability that at round $t$ AI offers a product $a$ is $p_A^t$. $C$ is a positive number that controls the speed of learning of AI.

The other aspect that AI decides is the price for the product. This choice, high(h) or low(l), depends also on the level of concordance between the type of product offered and the probability that the algorithm assigns to each type. That is, the probability of high price when $p_A^t > \frac{1}{2}$ is denoted by $P_A^H$ is equal to $P(h|p_A^t > \frac{1}{2}) = p_A^t$; and the probability of low price when $p_A^t > \frac{1}{2}$ is $P_A^L = P(l|p_A^t > \frac{1}{2}) = 1 - p_A^t$. If $p_A^t = \frac{1}{2}$, then $P(h) = P(l) = \frac{1}{2}$.

We are ready to formulate the steady state of the process of updating the belief of A-dominant type. Note that a strategy is defined by elements $\delta_{ij}^{kl}$, such that $i$ denotes the product $a$ or $b$; $j$ denotes the price of the product, $h$ or $l$; $k$ denotes the role (temporal preference) of the consumer, either $A$ or $B$; and $l$ denotes the consumer's willingness to pay (WTP) for the product, $H$ or $L$ . If $\delta_{ij}^{kl} = 1$, it means that the consumer buys product $i$ at price $j$ when her type is $k$ and her WTP is $l$. On the other hand, if $\delta_{ij}^{kl} = 0$, the consumer does not buy the product. Let us denote by $p_A^\infty$ the long run probability of being type $A$ learnt by AI and $p_i^j$ is the probability of having $A$ role $i$ being $j$-dominant; and $p_{k|i}^j$ means the probability of having $k$ WTP being $j$-dominant for a given role $i$. This probability must be the fixed point of the following condition:

$$p^A_A p^A_{H|A}[\delta^{ah}_{AH} p^\infty_A p^\infty_A + \delta^{al}_{AH} p^\infty_A (1-p^\infty_A) + (1-\delta^{bh}_{AH})(1-p^\infty_A)(1-p^\infty_A) + (1-\delta^{bl}_{AH})(1-p^\infty_A)p^\infty_A]+$$
$$p^A_A p^A_{L|A}[\delta^{ah}_{AL} p^\infty_A p^\infty_A + \delta^{al}_{AL} p^\infty_A (1-p^\infty_A) + (1-\delta^{bh}_{AL})(1-p^\infty_A)(1-p^\infty_A) + (1-\delta^{bl}_{AL})(1-p^\infty_A)p^\infty_A]+$$
$$p^A_B p^A_{H|B}[\delta^{ah}_{BH} p^\infty_A p^\infty_A + \delta^{al}_{BH} p^\infty_A (1-p^\infty_A) + (1-\delta^{bh}_{BH})(1-p^\infty_A)(1-p^\infty_A) + (1-\delta^{bl}_{BH})(1-p^\infty_A)p^\infty_A]+$$
$$p^A_B p^A_{L|B}[\delta^{ah}_{BL} p^\infty_A p^\infty_A + \delta^{al}_{BL} p^\infty_A (1-p^\infty_A) + (1-\delta^{bh}_{BL})(1-p^\infty_A)(1-p^\infty_A) + (1-\delta^{bl}_{BL})(1-p^\infty_A)p^\infty_A] = p^\infty_A$$

Given the solution of the above expression we can also compute the payoff per stage given the above probability. Formally:

$$E(p^\infty_A) =$$
$$p^A_A p^A_{H|A}[\delta^{ah}_{AH} p^\infty_A p^\infty_A u^{ah}_{AH} + \delta^{al}_{AH} p^\infty_A (1-p^\infty_A) u^{al}_{AH} + \delta^{bh}_{AH}(1-p^\infty_A)(1-p^\infty_A) u^{bh}_{AH} + \delta^{bl}_{AH}(1-p^\infty_A)p^\infty_A u^{bl}_{AH}]+$$
$$p^A_A p^A_{L|A}[\delta^{ah}_{AL} p^\infty_A p^\infty_A u^{ah}_{AL} + \delta^{al}_{AL} p^\infty_A (1-p^\infty_A) u^{al}_{AL} + \delta^{bh}_{AL}(1-p^\infty_A)(1-p^\infty_A) u^{bh}_{AL} + \delta^{bl}_{AL}(1-p^\infty_A)p^\infty_A u^{bl}_{AL}]+$$
$$p^A_B p^A_{H|B}[\delta^{ah}_{BH} p^\infty_A p^\infty_A u^{ah}_{BH} + \delta^{al}_{BH} p^\infty_A (1-p^\infty_A) u^{al}_{BH} + \delta^{bh}_{BH}(1-p^\infty_A)(1-p^\infty_A) u^{bh}_{BH} + \delta^{bl}_{BH}(1-p^\infty_A)p^\infty_A u^{bl}_{BH}]+$$
$$p^A_B p^A_{L|B}[\delta^{ah}_{BL} p^\infty_A p^\infty_A u^{ah}_{BL} + \delta^{al}_{BL} p^\infty_A (1-p^\infty_A) u^{al}_{BL} + \delta^{bh}_{BL}(1-p^\infty_A)(1-p^\infty_A) u_{BL} + \delta^{bl}_{BL}(1-p^\infty_A)p^\infty_A u^{bl}_{BL}]$$

Note that our payoff matrix is:

Table 10: Payoffs for consumer type A-dominant.

| | | Role A | | Role B | |
|---|---|---|---|---|---|
| | | High WTP | Low WTP | High WTP | Low WTP |
| Item a | high price | 0.20 | 0.02 | -0.20 | -2.00 |
| | low price | 2.00 | 0.20 | -0.02 | -0.20 |
| Item b | high price | -0.20 | -2.00 | 0.20 | 0.20 |
| | low price | -0.02 | -0.20 | 2.00 | 0.02 |

Notice also that the strategy "purchase when role and product coincides" has the following form (see Table 2):

Table 11: Strategy "purchase whenever the quoted price is less than or equal to the WTP"

| | AH | AL | BH | BL |
|---|---|---|---|---|
| ah | 1 | 1 | 0 | 0 |
| al | 1 | 1 | 0 | 0 |
| bh | 0 | 0 | 1 | 1 |
| bl | 0 | 0 | 1 | 1 |

Note also that the strategy "purchase only when you stand to make the largest possible profit" has the following form (see Table 3):

Table 12: Strategy "purchase only when you stand to make the largest possible profit"

| | AH | AL | BH | BL |
|---|---|---|---|---|
| ah | 0 | 0 | 0 | 0 |
| al | 1 | 0 | 0 | 0 |
| bh | 0 | 0 | 0 | 0 |
| bl | 0 | 0 | 1 | 0 |

Note that Tables 4, 1 and 3 have been reproduced here, as Tables 10, 11 and 12, to make the Appendix self-contained.

By considering the consumer's behaviour, we can determine the steady state probability $p_A^\infty$, which represents AI's belief. Specifically, taking the corresponding set of $\delta_{ij}^{kl}$ when the consumer follows a strategy purchase whenever the quoted price is less than or equal to the WTP, the asymptotic AI belief equals 0.75, and the corresponding average payment per round converges to 0.3148437. However, if the consumer adopts the strategy equal to that from strategy purchase only when she stands to make the largest possible profit, then the asymptotic AI belief drops to 0.5615528, and the corresponding average payment per round becomes 0.3077641 in the long run. In our case, if consumers follow the former strategy, they reveal their real type, whereas following the latter strategy results in little to no revelation. Next, Figure 11 shows this convergence for both cases.

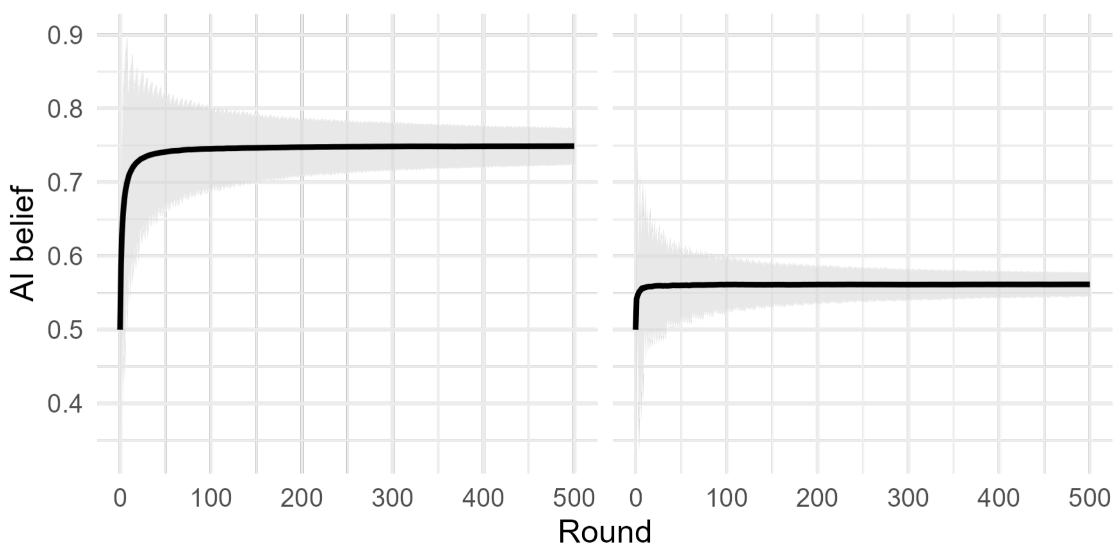


Figure 11: Convergence of AI beliefs when shoppers follow strategies, purchase whenever the quoted price is less than or equal to the WTP, (left panel), and purchase only when they stand to make the largest possible profit (right panel)

## 6.2 Appendix B. Analysis of sequence effects

In this section, we analyze whether there are learning effects that confound the second block's interpretation in some treatments. We run a regression of private information disclosed in Block 2 controlling for how participants performed in Block 1 to see if the difference in Block 2 can be fully attributed to the new treatment conditions. We find that the level of disclosure of private information in Block 2 does not depend on previous disclosing behavior but only on the new conditions.

| **Variable** | **Estimate** | **Std. Error** | **t value** | **p-value** |
|---|---|---|---|---|
| (Intercept) | 69.7301 | 7.5996 | 9.176 | $< 0.00001$ *** |
| AI Belief | 0.0551 | 0.0842 | 0.655 | 0.5134 |
| Awareness | 1.1994 | 6.0430 | 0.198 | 0.8429 |
| Info | -9.9775 | 6.0565 | -1.647 | 0.1011 |
| APP | -16.8552 | 6.0518 | -2.785 | 0.0059 ** |
| *Signif. codes:* 0 '***' 0.001 '**' 0.01 '*' 0.05 '+' 0.1 ' ' 1 | | | | |
| Multiple R-squared: 0.0585, Adjusted R-squared: 0.0392 | | | | |

Table 13: Private information disclosure in Block 2. OLS