Experimental Economics (2025), 1–22
doi:10.1017/eec.2025.10030

economic science association

CAMBRIDGE UNIVERSITY PRESS

ORIGINAL PAPER

Strategic information disclosure to classification algorithms: an experiment

Jeanne Hagenbach[1] and Aurélien Salas[2]

[1]CNRS, Sciences Po, WZB, CEPR, CESifo, Paris, France
[2]Sciences Po, Paris, France
Corresponding author: Jeanne Hagenbach; Email: jeanne.hagenbach@sciencespo.fr

(Received 28 January 2025; revised 26 June 2025; accepted 21 August 2025)

Abstract

We experimentally study how individuals strategically disclose multidimensional information to a Naive Bayes algorithm trained to guess their characteristics. Subjects' objective is to minimize the algorithm's accuracy in guessing a target characteristic. We vary what participants know about the algorithm's functioning and how obvious are the correlations between the target and other characteristics. Optimal disclosure strategies rely on subjects identifying whether the combination of their characteristics is common or not. Information about the algorithm functioning makes subjects identify correlations they otherwise do not see but also overthink. Overall, this information decreases the frequency of optimal disclosure strategies.

Keywords: Classification algorithms; data management; experiments; strategic disclosure

JEL Codes: C91; D89; D91; M38

1. Introduction

Since 2018, the *General Data Protection Regulation* has imposed obligations on any organization that collects data related to people in the European Union. Two of the principles identified as key to ensure data privacy and security are *algorithmic transparency* and *user control*. The first principle prescribes that individuals must be informed in a "concise, transparent, intelligible and easily accessible" way about how their data is processed (see Art 12-14 GDPR). The second principle provides individuals with some control over their personal data in the precise sense of a "right to object, at any time, to processing of [this] data" (see Art. 21 GDPR). Underlying these principles is the assumption that informed individuals who have the possibility to manage their data will effectively do it in their best interest.

We propose an experiment to test this assumption and examine what helps subjects manage their data. Subjects face a classification algorithm trained on other individuals' data to guess their personal attributes. Classification algorithms are nowadays prevalent. They segment people into categories which predict who they are, what they will do or like.[1] We implement a stylized, simplified version of these situations in which individuals' objective is clearly defined – they must prevent the algorithm from guessing one specific personal attribute - and their task consists in strategically disclosing or

[1]For example, classification algorithms are used to target advertisements and recommend content (Basu et al., 1998), categorize job applicants (Pal et al., 2022) or group individuals by levels of risk (Rawat et al., 2021).

hiding only six attributes. We show that less than 40% of individuals' data management decisions are optimal and, while information helps subjects understand some aspects of the problem, it does not always help them play more optimally.

Our pre-registered experiment is made up of three parts. In the first part, subjects answer six binary questions about themselves. This part allows us to obtain a set of characteristics for every individual: gender, marital status, children, time spent weekly listening to music, and preferences about ice cream flavor and nuclear power. In the second part, subjects play against an algorithm which does not know their individual characteristics but has been trained to guess them. In each round of play, the subjects' goal is to prevent the algorithm from guessing the answer they gave to one of the six questions of Part 1, the *target question* for that round. To do so, subjects must decide, for each answer they gave in Part 1, whether to disclose or hide it to the algorithm (no lie possible). Experimental payments represent a trade-off commonly faced by individuals disclosing personal data to algorithms. On the one hand, hiding characteristics is costly, translating the idea that it takes time and effort to prevent algorithms from accessing personal data. On the other hand, hiding characteristics reduces the information that the algorithm can use to guess subjects' characteristics. In the experiment, subjects' payment is inversely proportional to the accuracy with which the algorithm guesses their answers to the target questions. In the third part of the experiment, we ask subjects to report the correlations they see between the questions of Part 1.

The algorithm we use is a Naive Bayes Classifier. For a given target question, this algorithm guesses the subject's answer according to the following main principles. First, it only uses the answers disclosed by the subject and does not deduce anything from the hidden (missing) characteristics. Second, to guess the probability of a given answer to the target question conditional on a set of disclosed answers, the algorithm uses Bayes' rule with the 'naive' assumption that the subject's characteristics are mutually independent conditional on the characteristic it is trying to guess. This assumption ensures that every disclosed answer independently contributes to the algorithm guess, thereby simplifying subjects' task. Third, to compute its guess, our algorithm uses the prior and conditional probabilities of the different characteristics which are the frequencies of these characteristics in a population of around 500 individuals. These individuals participated in a pre-study which generated the training data for the algorithm. In short, our algorithm uses existing correlations between the characteristics to guess subjects' answers based on the partial information they disclose.

In order to 'game' the algorithm at the lowest cost, subjects need to understand how it functions and, once they know it uses correlations, to properly identify these correlations. Our experimental treatments involve variations along these two dimensions. First, we vary, between subjects, the information given about the functioning of the algorithm: in *Control*, subjects are simply told that the algorithm uses the answers they disclose to guess their answer to the target question; in *Info*, subjects are additionally told that the algorithm uses correlations between answers to deduce theirs, and that it has been trained on the answers of 500 individuals to identify these correlations. Second, subjects play with four different target questions. This allows us to consider both strong and absent correlations between target questions and other questions, as well as vary how obvious the correlations are. Precisely, we consider the following four targets: two questions, abbreviated ICE and MUS for favorite ice cream and time spent listening to music, whose answers are not correlated to any other answers given in Part 1; the question about marital status, abbreviated MAR, whose answer is highly and obviously correlated to the answer about having children; the question about being favorable to nuclear power, abbreviated NUC, whose answer is correlated to gender.

We examine subjects' behavior by considering the number of answers they hide from the algorithm and the frequency with which they use optimal disclosure strategies. Given our experimental payoffs and training data, optimal strategies, which we characterize for every subject and target question, turn out to be relatively intuitive. When the target question is uncorrelated (ICE and MUS), it is optimal for subjects to hide only the answer to the target question itself. When the target question is correlated (MAR and NUC), the optimal strategy depends on whether subjects are *common* or

https://doi.org/10.1017/eec.2025.10030 Published online by Cambridge University Press

uncommon, that is, on whether they answered like the majority of individuals in the pre-study or not. Common subjects should hide the answer to the target question and the answer to the question which is most correlated to that target. For example, if subjects do not want the algorithm to guess they are married, they should hide that the fact they have children. In contrast, uncommon subjects should hide only the answer to the target question. For example, non-married subjects with children should disclose having children to mislead the algorithm into guessing they are married. Our paper is the first to shed light on the distinction between common and uncommon subjects, which is key to play optimally against the algorithm.[2]

The first results relate to the aggregate effects of our two experimental variations. First, pooling all target questions, the frequency with which subjects play optimal strategies is lower in *Info* than in *Control*. Contrary to what we had hypothesized and pre-registered, subjects disclose less optimally when they have more information about the functioning of the algorithm. Overall, information pushes subjects to over-think: they see correlations which do not exist and hide more answers than what is optimal. As we will see in the next paragraph, this general observation hides important differences between the target questions. Second, conditional on the treatment being *Control* or *Info*, the frequency of optimal strategies is lowest when the target question is NUC. This confirms the pre-registered hypothesis according to which subjects play better against the algorithm when they understand well how their answer to the target question correlates to other answers.

Our main result is that the effect of information, not beneficial for subjects overall, varies drastically with the initial level of knowledge that subjects have about existing correlations. When the target questions are ICE and MUS, subjects understand well the absence of correlations and that it is optimal to hide the answer to the target question only. The *Info* treatment pushes subjects to search for nonexistent correlations and play sub-optimally. When the target question is MAR, a large majority of the subjects identify well that this question is correlated to the question about children, and the *Info* treatment does not significantly change the way they play. When the target question is NUC, the *Info* treatment helps a significant share of subjects find the correlation to gender. It however does not help them to understand the direction of this correlation, which is needed to play optimally.

Finally, we try to examine in more details the logic behind subjects' disclosure choices (at least for subjects who, in the first place, hide the answer to the target question itself). We show that a majority of subjects choose disclosure strategies that are consistent with the correlations they have identified, and that the level of consistency is similar in *Control* and *Info*. However, as explained in the previous paragraph, information changes the correlations identified by the subjects, sometimes in the direction of a lower accuracy. This explains part of the overall negative effect of *Info* on the optimality of players' strategies. For another part, subjects who understand correlations properly still need to be sophisticated enough to play differently when being common or uncommmon.

Related literature. First, our paper is related to theoretical works studying situations in which agents input private data into systems which generate payoff-relevant outcomes for them. These works span computer science, statistics and economics.

In computer science and statistics, the focus is on building algorithms that are robust to strategic manipulation of their data by the agents. In Meir et al. (2012), experts with personal interests provide training data to classification algorithms. In a seminal article, Hardt et al. 2016 consider individuals who can manipulate their attributes at some cost to obtain better classification outcomes. For certain instances of these problems, the authors propose algorithms which achieve minimal classification errors.[3] We study individuals' strategies for a fixed algorithm rather than adapt the algorithm to these strategies.

[2] A similar distinction between gender-stereotypical and non-gender-stereotypical personal attributes appear in Slokom et al. (2021). This distinction is used to design recommendation systems which keep gender private.

[3] Extensions to this work include Kleinberg and Raghavan (2020) which examine individuals' efforts to manipulate their attributes, Krishnaswamy et al. 2021 which considers agents withholding information instead of lying, and Hu et al. 2019 which consider heterogeneous gaming abilities.

https://doi.org/10.1017/eec.2025.10030 Published online by Cambridge University Press

In economics, Frankel and Kartik (2022), Perez-Richet and Skreta (2022) and Ball 2024 consider the problem of a designer who commits to a mechanism or a test which determines allocations or scores as a function of agents' reports. These works show how to use reported information in a way that induces more truthful revelation from agents who want higher allocations or scores. Björkegren et al. (2020) is close in spirit to these papers in that it designs an algorithm that is robust to manipulation by agents, and tests it in a large field experiment in Kenya. In a different type of work, Eliaz and Spiegler (2019), Eliaz and Spiegler 2022 consider a statistician using a penalized regression model to determine the best action for an agent. The statistician and the agent have aligned interests, but sampling errors and penalties for including variables in the model can create incentives for the agent to misreport his characteristics. In most of the above-mentioned theoretical works, agents, at least some of them, are assumed to be sophisticated enough to adjust to the mechanisms or the models they face. We evaluate this sophistication experimentally. In a closely-related model, Miklós-Thal et al. (2024) consider agents who disclose multi-dimensional data to a firm. The firm infers hidden information from disclosed information using correlations deduced from gathering "users" data over time. In the long term, when users are aware of these correlations, they either disclose all information or become digital hermits who hide all information.

Second, our paper is linked to a large experimental literature in economics and psychology which study how individuals' attitude towards privacy affects online information disclosure. In comprehensive surveys, Acquisti et al. (2017) and Acquisti et al. 2020 discuss various factors at the origin of the *privacy paradox*, the frequently-observed disconnection between stated privacy preferences and actual behavior: individuals prioritize immediate rewards over long-term privacy (Acquisti, 2004), individuals disclose more sensitive information when they perceive others are doing so (Acquisti et al., 2012), individuals stick with default revelation options leading to less (John et al., 2011), and so on. Our results show that managing personal data is challenging for subjects even when abstracting away from privacy concerns.

Bó et al. (2023) report an experiment closely related to ours. They study how users manipulate their responses to a questionnaire in order to achieve favorable pricing in a price discrimination setting. They show that users effectively manipulate their answers only when the link between these answers and the proposed price is direct and obvious. In their study, subjects can lie whereas we focus on hard information disclosure. The algorithm used in Bó et al. (2023) is an OLS regression which estimates subjects' willingness-to-pay from their answers; our algorithm is a Naive Bayes classifier trained to guess personal attributes, which could be used subsequently for various purposes. Using different approaches, both papers suggest that transparency and user control still lead to sub-optimal disclosure decisions. More broadly, our paper relates to a body of work studying why individuals are averse to rely on algorithms for some tasks typically done by humans, even if algorithms often perform better (see Dietvorst et al. (2015), Dietvorst et al. (2018), Castelo et al. (2019), or Jussupow et al. (2020)). Within this literature, Dargnies et al. 2024 focuses on hiring algorithms and study the effect of transparency on their adoption. We study how transparency affects subjects' strategic communication with algorithms.

2. Experimental design

We describe the overall structure of the experiment before giving details about the treatments. To develop and train the algorithm against which subjects play in the experiment, we collected data in a pre-study.

2.1. Pre-study

The pre-study involved 505 Prolific participants (fluent in English and based in the USA) who completed a simple task: They answered 30 binary questions about themselves.[4] Subjects were paid

[4]Only the question about gender was not binary. The pre-study questionnaire is given in the Online Appendix.

https://doi.org/10.1017/eec.2025.10030 Published online by Cambridge University Press

a fixed amount of 60 pence for filling out that questionnaire, which took on average 2 minutes 51 seconds. We had explained to the subjects that there were no right or wrong answers and that they should answer honestly, so we consider their answers as truthful.

2.2. Main experiment

Our experiment is made up of three parts. The instructions for each part are given to subjects along the way. Subjects complete each part without knowing what the next part is made of. Subjects can earn money in each part as detailed below. The complete instructions are given in section 3 of the Online Appendix.

2.2.1. Part 1

In the first part, each subject completes a short questionnaire consisting of six questions about demographics and preferences. The questions are presented in one of three random orders, and, as in the pre-study, subjects are asked to answer honestly. For completing the questionnaire, subjects receive a fixed payment of £1.2. The six questions and possible answers are given below. We explain in section A.4 of the Appendix how we selected these questions from the questionnaire and data of the pre-study. In the paper, we refer to each question by using the three letters which appear below, before each question.

CHI - Do you have children? *Yes / No*

GEN - What gender are you currently? *Male / Female / Non-Binary.*[5]

MAR - Are you married or in a domestic partnership? *Yes / No*

MUS - How much time do you spend listening to music per week? *3 hours or less / More than 3 hours*

ICE - Which flavor of ice cream do you prefer? *Chocolate / Vanilla*

NUC - Are you in favor of the use of nuclear power? *Yes / No*

2.2.2. Part 2

In the second part of the experiment, subjects play four rounds of a game against an algorithm. The general idea of this game is as follows: The algorithm does not know the subjects' answers to the questionnaire completed in Part 1 but it is trained to guess them. In every round, the subjects' objective is to prevent the algorithm from guessing their answer to one specific question asked in Part 1. We refer to this question as the *target question*. Each round of the game proceeds in three steps.

- First, we tell the subject which question is the target question and remind him/her the answer he/she gave in Part 1.

- Second, the subject must decide, for each of the six answers he/she gave in Part 1 (including the answer to the target question), whether or not he/she wants to disclose it to the algorithm. We do not offer subjects the possibility to manipulate the answer they gave, only to hide it from the algorithm (no lies are possible). Figure 1 is a screenshot of the interface subjects used to make these choices.

- Third, once the subject made his/her six disclosure decisions, the algorithm uses the disclosed answers to compute a probability for each possible answer to the target question (the algorithm does not deduce anything from undisclosed answers). For example, if the target question is "Do you have children?", the algorithm computes the probability that the answer of the subject was "yes" and the complementary probability that the answer of the subject was "No". The probability that is computed for *the answer effectively given by the subject* in Part 1 is called the *guess of the algorithm*.

[5]We did not have enough *Non-binary* participants in the pre-study to train the algorithm properly for these subjects. In the main experiment, the 12 subjects who answered *Non-Binary* to the gender question could play but were later dropped from the main analysis.

https://doi.org/10.1017/eec.2025.10030 Published online by Cambridge University Press

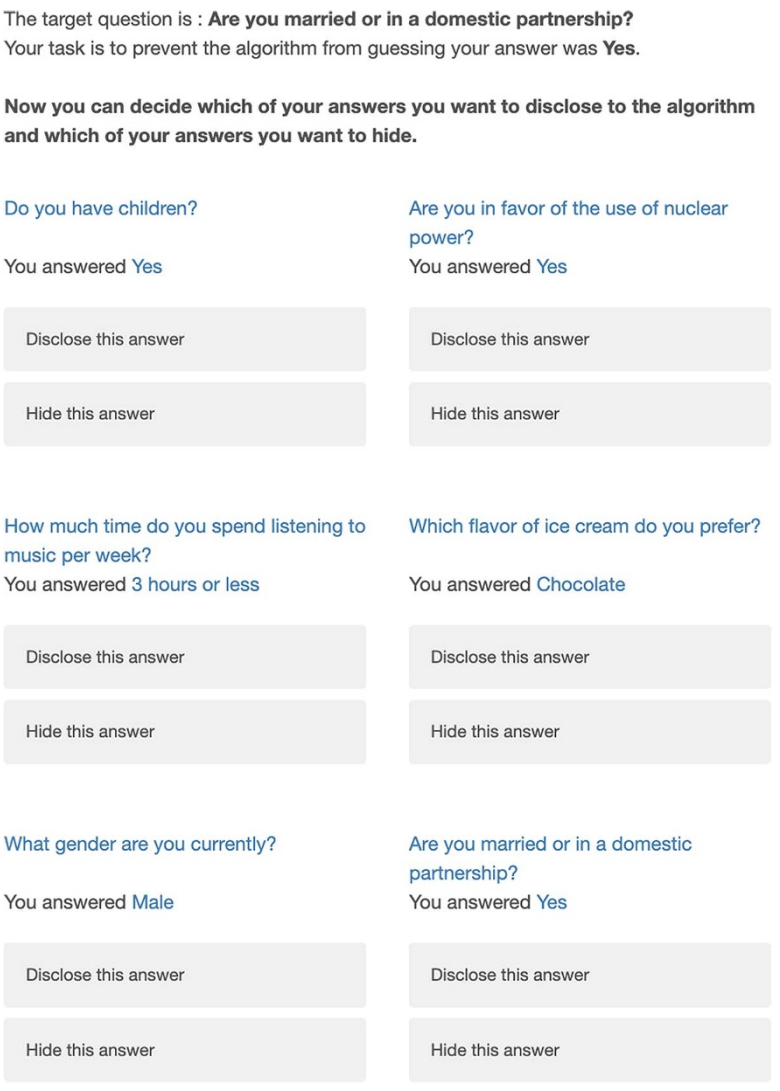

Fig. 1 A screen seen by subjects when they had to make their disclosure choices

We give the details of the subjects' payments later but the key trade-off is the following: hiding answers to the algorithm is costly to the subject but, if done strategically, can prevent the algorithm from making more accurate guesses.

2.2.3. How does the algorithm compute its guesses?

We now describe the environment more formally to explain how the algorithm computes its guesses in every round of the game. In the environment we consider, there are six binary random variables, $\tilde{x}_1$ to $\tilde{x}_6$, each corresponding to a question asked in Part 1. Every subject is characterized by the realizations of these variables, that is, by the set of six answers he/she gave in Part 1, $A \equiv \{x_1, x_2, x_3, x_4, x_5, x_6\}$, and discloses a subset of these answers, $D \subseteq A$, to the algorithm.

The first property of the algorithm we implement is that it uses only disclosed answers to make its guesses, in the sense that it does not make any inferences from hidden answers. Next, the algorithm we implement is the Naive Bayes Algorithm: when the target question is j and the subject discloses

https://doi.org/10.1017/eec.2025.10030 Published online by Cambridge University Press

D, the guess of the algorithm corresponds to $g_D \equiv P(x_j|D)$ which is computed using Bayes' rule with the "naïve" assumption that all variables in $\{\tilde{x}_i\}_{i\neq j}$ are mutually independent conditional on x_j. The assumption of conditional independence simplifies considerably. the relationship between disclosed answers and the guess. Because each disclosed answer contributes independently to the guess, subjects can think about the effect of disclosing each answer in isolation.[6]

Formally, when $D \neq \emptyset$, the algorithm's guess is given by:

$$g_D^j \equiv P(x_j|D) = \frac{P(x_j)\prod_{x_i\in D} P(x_i|x_j)}{P(D)}. \tag{1}$$

where

$$P(D) = Pr(x_j)\prod_{x_i\in D} P(x_i|x_j) + Pr(\neg x_j)\prod_{x_i\in D} P(x_i|\neg x_j).$$

To compute the guesses, according to (1), the algorithm only uses the prior probabilities $P(x_j)$ of target questions j and the conditional probabilities $P(x_i|x_j)$ for every i and target j. It computes these probabilities using frequencies from the pre-study dataset. Precisely, $P(x_j)$ correspond to the frequency of answer x_j in the pre-study dataset, and $P(x_i|x_j)$ to the frequency with which the answer x_i occurs in conjunction with x_j divided by the frequency of x_j. In the particular case in which the subject discloses the answer to the target question j itself, then (1) leads to $g_D^j = 1$.[7] If the subject does not disclose any answer, $D = \emptyset$, equation (1) is not defined. The guess of the algorithm then simply corresponds to the prior probability of answer x_j, $P(x_j)$, given by the frequency of x_j in the pre-study data.

The Naive Bayes Algorithm is widely used in practice.[8] According to Wu et al. (2007), Naive Bayes ranks among the top 10 algorithms used in both the industry and the academic world for analyzing large datasets. Its applications range from medical classification tasks, such as predicting cancer progression in patients (Kamel et al., 2019), to everyday use like spam filtering in software such as Apache SpamAssassin and Mozilla Thunderbird. Naive Bayes algorithms also perform well for recommendation tasks. For instance, Wang and Tan 2011 shows that an improved version of the Naive Bayes algorithm performs better than the Amazon recommendation algorithm, and Sahu et al. 2017 found the Naive Bayes approach to be the most precise for movie recommendations. In addition, Pronk et al. 2007 and Valdiviezo-Diaz et al. (2019) argue, respectively, in favor of the Naive Bayes Algorithm because it is relatively simple to use and to explain to individuals.

2.2.4. Subjects' payment in part 2

In each round of the game against the algorithm, the payoffs are as follows. The subject starts each round with an endowment of £3.2. This endowment is reduced in two ways: (1) For each answer that the subject decides to hide from the algorithm, the endowment is reduced by 20 pence. (2) At the end of the round, the endowment is reduced by two times the guess (between 0 and 1) of the algorithm. We remind that this guess corresponds to the probability, computed by the algorithm, that the subject's answer was the one truly given in Part 1. With such payoffs, hiding answers is costly but reduces the information available to the algorithm to guess the subject's answers. We will later show that reducing this information can have ambiguous effects on how accurate the algorithm guess is, and derive subjects' optimal disclosure strategies.

[6] If the algorithm did not assume that all variables $\{\tilde{x}_i\}_{i\neq}$ are independent conditional on x_j, it would be hard for subjects to evaluate the effect of disclosing an answer on the guess. They would have to think about the direct effect of this answer on the guess, but also about the indirect effect on other answers which also affect the guess.

[7] This is true in theory. In practice, algorithms need to avoid break-downs linked to zero probabilities, so they apply smoothing methods to their computations. Our algorithm delivers a guess of at least 0.983 for the cases in which subjects disclose the answer to the target question.

[8] We use the *BernouilliNB* code in the Python *sklearn* package. For details, see section 4 of the Online Appendix.

https://doi.org/10.1017/eec.2025.10030 Published online by Cambridge University Press

Before starting the four rounds of game, subjects need to answer correctly some comprehension questions. Once a round is over, subjects move to the next round without getting any feedback about the guess of the algorithm. Each round corresponds to a different target question, and the order of the four rounds/target questions is randomized at the subject level. Rounds are independent in the sense that the answers disclosed in one round by the subject cannot be used by the algorithm in the next rounds. One of the four rounds is picked at random for the payment of Part 2 of the experiment.

2.2.5. Part 3

Part 3 consists of a questionnaire whose goal is to get a sense of the correlations that subjects see between the answers to the six questions of Part 1. For each of the four target questions, we ask subjects to imagine they would have to guess someone's answer to that target question. Then we ask, if they could see this person's answer to one other question, which they think would be most useful. To capture the possibility that subjects see no correlations between the target question and the other questions, we offer subjects the option to answer "none of the questions would help me much to make that guess". For every correct answer given by the subjects in the Part 3 questionnaire, that is, when they can identify the most correlated question or rightly identify that the target question is not correlated to any other question, they get 10 pence. Finally note that, in Part 3, we elicit whether subjects see correlations but do not ask them the direction of these correlations.

At the very end of the experiment, subjects are asked about their age and experience with algorithms, Internet and statistics.

2.2.6. Implementation

The experiment was run on Prolific and involved 970 subjects (fluent in English and based in the USA).[9] The experiment took, on average, 8 minutes and 43 sec. (sd 5 minutes and 11 sec.) and subjects earned an average of £2.99 (sd 50 pence). (The pre-registration - reference #128706 on Aspredicted - included an additional treatment, presented and briefly analyzed in section 2 of the Online Appendix.)

2.3. Experimental treatments

Our objective is to understand what affects subjects' ability to "game" the algorithm, that is, to prevent the algorithm from guessing their answers with a high probability. Subjects may fail to do so for at least two reasons. One reason is that they do not know how the algorithm functions and, in particular, that it uses correlations between questions to make guesses. Another reason is that, even if they understand that the algorithm uses correlations to make guesses, they do not identify which questions are correlated to each other, and which are not correlated to any other. We design two dimensions of treatments along these two lines: One dimension varies the information we give to subjects about the functioning of the algorithm; the other dimension varies how easy it is for subjects to understand the correlations or the absence of correlations.

2.3.1. Variation 1: information about the algorithm

Subjects are randomly assigned to the *Control* or to the *Info* treatment (between subjects implementation). In the beginning of Part 2, we explain to the subjects the game they will play against the algorithm and, in particular, give them the following information:

- In the *Control* treatment, subjects read: *In every round, you will have to decide, for each answer you gave in Part 1, whether you want to disclose it or hide it to the algorithm. The algorithm will use the answers you disclose to deduce your answer to the target question.*

[9] As explained earlier, out of 982 in total, we had to drop the 12 subjects who answered *Non-Binary* to GEN.

https://doi.org/10.1017/eec.2025.10030 Published online by Cambridge University Press

– In the *Info* treatment, subjects read the same sentences as in the *Control* treatment but we add the following text: *To make this deduction, the algorithm has been trained on about 500 subjects, who previously completed the same questionnaire as the one you completed in Part 1. The algorithm uses their answers to identify correlations between answers. For example, it can identify whether women are more or less likely than men to listen to more than three hours of music per week.*

2.3.2. Variation 2: correlations between target questions

Every subject plays four rounds of the game against the algorithm. In every round, the target question is different. We selected target questions which were not correlated to each other and with different levels of correlation to other questions.

We use ICE, MUS, MAR and NUC as target questions.[10] In the pre-study dataset, the correlation between ICE and MUS and any other question is lower than 0.10. We refer to ICE and MUS as *uncorrelated target questions*. In the pre-study dataset, the answer to MAR is correlated to the answer to CHI (Pearson correlation coefficient is 0.47) and, more precisely, subjects who are married are also more likely to have children (and vice versa). The answer to MAR is not correlated to the answer to any question other than CHI. Finally, NUC is correlated to GEN (Pearson correlation coefficient is 0.29) and, more precisely, male subjects are more likely to be in favor of the use of nuclear power (and vice versa).[11] Again, the answer to NUC is not correlated to the answer to any question other than GEN. We refer to MAR and NUC as *correlated target questions*.

Finally, we assume that the correlation between MAR and CHI is easier to identify for subjects than between NUC and GEN. We also assume that the absence of correlation of MUS and ICE with any other question is easier to see than the correlation between NUC and GEN. At the end of section 3.1, we give arguments supporting these assumptions.

2.4. Optimal disclosure strategies

In this subsection, we derive the subjects' optimal disclosure strategies before discussing the generality of the theoretical predictions established for our experimental setting.

To find optimal disclosure strategies, we must find, for every subject, the largest disclosure set (hiding is costly) which prevents the algorithm from making too accurate guesses. As mentioned above, a subject is characterized by the six answers he/she gave in Part 1, $A = \{x_1, x_2, x_3, x_4, x_5, x_6\}$. $D \subseteq A$ is the set of answers disclosed by the subject. When the target is j, the guess g_D^j of the algorithm is given by (1) if the set of disclosed answers is $D \neq \emptyset$ and by $P(x_j)$ if $D = \emptyset$. Given experimental payoffs, the subject's objective is the following:

$$\max_{D \subseteq A} \quad -2g_D^j - 0.2 * |A \setminus D|$$

where $|A \setminus D|$ corresponds to the number of answers hidden by the subject.

To establish results about optimal strategies for all subjects, we need to consider all possible sets of answers A. For a given A, we then need to compare the subjects' payoffs for all possible disclosure strategies. For that, we design a procedure which compares disclosure strategies two by two for a given A, and then repeats this exercise for all possible A. The proofs of the two following propositions are given in section A.5 of the Appendix.

We start by establishing a rather intuitive result, namely that it is always beneficial for a subject to hide the answer to the target question itself.

Proposition 1. *In the game against the algorithm, it is always strictly beneficial for the subjects to hide the answer to the target question.*

[10] Section A.4 of the Appendix explains how we selected these questions.

[11] This result is not specific to our sample. Solomon et al. (1989) and Kennedy et al. (2023) report a similar link between being a man and the acceptance of nuclear power.

https://doi.org/10.1017/eec.2025.10030 Published online by Cambridge University Press

For the uncorrelated target questions, ICE and MUS, hiding only the target answer is always optimal. Intuitively, since the target questions are uncorrelated, hiding additional answers will have only a negligible effect on the guess of the algorithm while costing 20 pence. For the correlated target questions, MAR and NUC, the guess of the algorithm is strongly determined by the answer to the question that is correlated to the target (if disclosed), respectively CHI and GEN. For subjects who answered these questions like the majority of subjects in the pre-study, disclosing these answers help the algorithm make a better guess about their answer to the target question. For subjects who answered differently from the majority of subjects in the pre-study, these answers mislead the algorithm about their answer to the target question. For each target question, we respectively define *common* and *uncommon* subjects as follows.

Definition 1. Let the target question be MAR. A *common* subject either answered *Yes* to both MAR and CHI, or answered *No* to both MAR and CHI. An *uncommon* subject either answered *Yes* to MAR and *No* to CHI, or answered *No* to MAR and *Yes* to CHI.

Definition 2. Let the target question be NUC. A *common* subject either answered *Yes* to NUC and *Male* to GEN, or answered *No* to NUC and *Female* to GEN. An *uncommon* subject either answered *Yes* to NUC and *Female* to GEN, or answered *No* to NUC and *Male* to GEN.

We now can give the optimal strategies for all target questions and types of subjects.

Proposition 2.

(a) When the target question is uncorrelated (ICE or MUS), it is optimal for every subject to hide only the answer to the target question.
(b) When the target question is correlated (MAR or NUC), it is optimal for every common subject to hide the answer to the target question and the answer to its correlated question (resp. CHI or GEN).
(c) When the target question is correlated (MAR or NUC), it is optimal for every uncommon subject to hide only the answer to that target question.

The above propositions are established for the specific costs - cost to hide an answer, cost to see the answer to the target question accurately guessed - implemented in our experiment. Let us discuss the case in which subjects' payoff is, more generally, given by

$$-\alpha g_D^j - \beta * |A \setminus D|$$

with $\alpha \geq 0$ parameterizing the cost incurred by subjects when the algorithm guess becomes more accurate and $\beta \geq 0$ the cost incurred for every hidden answer. If α goes to zero, the cost to be identified is so small compared to the costs of hiding that subjects should disclose all answers. If β goes to zero, the cost of hiding is negligible: subjects' optimal strategy consists in hiding all answers which were also given by a majority of the subjects who gave the same answer as themselves to the target question, and disclosing all answers which were given by only a minority of the subjects who gave the same answer as themselves to the target question.[12] In that sense, the common versus uncommon distinction presented in Proposition 2 captures a general feature of optimal disclosure strategies. Our experiment, however, does not implement any of these two extreme cases, so we identify optimal disclosure strategies by comparing the cost to hide any question to the impact it has on the algorithm's

[12]Given equation (1), it is straightforward to show that, for all target question j, other question k and disclosed set D, $g_D^j - g_{D\setminus\{x_k\}}^j \geq 0$ (≤ 0, resp.) if and only if $P(x_k \mid x_j) \geq 0.5$ (≤ 0.5, resp.).

https://doi.org/10.1017/eec.2025.10030 Published online by Cambridge University Press

guess. Given our training dataset, all our Propositions hold as long as the ratio between β and α lies in $[0.053, 0.12]$.[13]

2.5. Hypotheses

The optimal strategies are relatively simple as they all consist in hiding only one or two answers. The objective of this paper is to study what affects players' ability to play these strategies. Clearly, to play optimally, subjects need to understand how the algorithm functions and, provided they understand it uses correlations, to identify these correlations.

The first treatment variation varies whether or not subjects were informed that the algorithm makes its guesses using correlations. Regarding this variation, we make the following, pre-registered, hypothesis:

Hypothesis 1. For every target question, subjects play the optimal strategy more often in the *Info* treatment than in the *Control* treatment.

The second treatment variation aims at examining how subjects play when correlations or absence of correlations between questions are more or less easy to identify. Regarding this variation, we make the following, pre-registered, hypothesis:

Hypothesis 2. Given a level of information about the functioning of the algorithm, subjects play the optimal strategy more often when the correlations or absence of correlations are easier to identify. Hence, subjects play the optimal strategy more often when the target question is MAR, ICE or MUS than when it is NUC.

3. Results

3.1. Description of the data

In our dataset, an observation corresponds to one of the four games played by one of the 970 subjects. We have 3880 observations in total, each consisting in a set of answers A given by the subject, a target question j and a set of disclosed answers D. For each observation, the previous propositions characterize the optimal disclosure strategy. The first three lines of Table 1 give the number of observations per treatment and target question. The bottom part of the table gives, for each target, the percentage of cases in which the optimal disclosure strategy is to hide only the answer to the target question. For MAR and NUC, this percentage corresponds to the fraction of uncommon subjects (defined only for correlated targets).

Regarding subjects' characteristics A, the answers to each of the six binary questions in Part 1 are well balanced: No answer is given by more than 62% of the subjects and no answer is given by less than 38% of the subjects. The proportion of each answer is not significantly different in *Control* and *Info*, except for slightly fewer married subjects in *Info*. In the subsequent analysis, one additional subjects' characteristic will prove relevant, namely whether or not subjects had already taken a course in statistics. This is the case for almost half of the subjects (46.49%) and highly correlated to educational attainment (Pearson correlation coef. of 0.48). Details about characteristics are given in section A.1 of the Appendix.

Regarding disclosure strategies, we start with a few general remarks before examining in detail how subjects play in the next sections. First, according to Proposition 1, subjects should always hide

[13] The bound 0.12 corresponds to the impact on the algorithm guess of hiding GEN when the target question is NUC and for the common subject for whom doing this has the smallest impact. The bound 0.053 corresponds to the impact on the algorithm guess of hiding CHI additionally to GEN when the target question is NUC and for the common subject for whom doing this has the largest impact. Both bounds appear in the proof of Proposition 2.

https://doi.org/10.1017/eec.2025.10030 Published online by Cambridge University Press

Table 1. Summary of data

	MUS	ICE	MAR	NUC	Total
Control	477	477	477	477	1908
Info	493	493	493	493	1972
Total	970	970	970	970	3880
Hiding the target only is optimal (in %)	100	100	25.36	35.26	65.15
Fraction of uncommon subjects (in %)	–	–	25.36	35.26	–

the answer to the target question. This result is intuitive for subjects who have understood the game and we use it to check whether they did: In 79.95% of the rounds, subjects indeed hide this answer from the algorithm;[14] 78.56% of subjects hide the answer to the target question in at least three of the four rounds they play. These statistics do not significantly differ for the *Info* and *Control* treatments, nor for the different target questions. Second, in every round of game, subjects decide whether to hide or disclose each of the six answers they gave in Part 1 which results in 64 possible strategies. Two such strategies can be considered as relatively "naive" in that they consist either in hiding all answers or in disclosing all answers; they are respectively used in 3.69% and in 10.34% of the cases. Since hiding is costly but disclosing the target question helps the algorithm too much, another "natural" strategy consists in hiding only the answer to the target question. This strategy, sometimes optimal, is used widely, namely in 34.23% of all cases.

The data also contain the answers given by subjects in Part 3 of the experiment. These answers indicate, for each target question, which other question is considered by the subject as most correlated to the target, if any. Half of the 3880 answers (50.59%) given in Part 3 are correct, that is, correspond to a case in which the subject identifies well the strongest correlation or the absence of correlation. In section A.2 of the Appendix, we summarize all answers given by subjects in Part 3. These answers support our assumption that the correlation between NUC and GEN is harder to identify for subjects than the correlation between MAR and CHI or the absence of correlation for ICE and MUS. For the uncorrelated targets ICE and MUS, the most common answer (respectively 50.21% and 46.08% of answers) is that these questions are correlated to no other question; about 80% of subjects answer that the MAR target is correlated to CHI; for the NUC target, the most common answer (38.35% of answers) is that it is correlated to no other question, which is incorrect.

3.2. Overall effect of information

In what follows, we analyze subjects' disclosure strategies by considering two main experimental outcomes: The frequency of optimal strategies and the number of hidden answers.[15] We start by examining the effect of the *Control* and *Info* treatments on these outcomes at the aggregate level, that is, by pooling all target questions.

Over all observations, subjects play the optimal strategy 33.97% of the time. This frequency equals 37.26% in *Control* against 30.78% in *Info*, which is significantly lower ($p < 0.001$). This means that, at the aggregate level, subjects play significantly less well when informed that the algorithm uses correlations to deduce their answers. This finding invalidates hypothesis 1 and is confirmed by the

[14]In the data we analyze, we keep the observations in which subjects disclose the answer to the target question. Our results are robust to dropping these observations and to dropping subjects who disclosed the answer to the target at least once over the four rounds they play.

[15]For several reasons, it is hard to use subjects' realized payoffs to evaluate subjects' ability to play the game. First, each subject's best possible payoff depends drastically on his/her specific answers to the initial questionnaire. In particular, common and uncommon players reach very different payoffs when playing optimally. To control for these differences, we could look at the difference between subjects' realized and best possible payoffs. Again, we can show that such a measure is problematic: The consequences on realized payoffs of any given strategic error are not the same for two people with different characteristics.

https://doi.org/10.1017/eec.2025.10030 Published online by Cambridge University Press

Table 2. Optimal strategies - all targets

	Optimal strategy		
	(1)	(2)	(3)
Info	−0.065***	−0.065***	−0.067***
	(0.021)	(0.021)	(0.021)
Round		0.019***	0.019***
		(0.006)	(0.006)
Stats			0.050**
			(0.021)
Female			0.007
			(0.021)
Age			−0.002**
			(0.001)
Constant	0.373***	0.326***	0.380***
	(0.016)	(0.021)	(0.043)
Observations	3880	3880	3880

Note: The table reports OLS coefficients (standard errors, clustered by subject, appear in parentheses). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

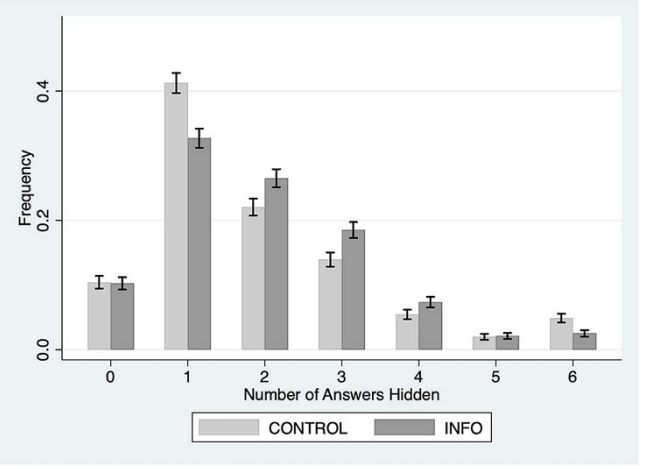

Fig. 2 Hiding 0 to 6 answers, per treatment and pooling all targets

regressions presented in Table 2. In this table, we examine the effect of the *Info* treatment dummy, the *Round* of play (ranging from 1 to 4), and subjects' demographics (gender, age and whether or not they took a course in *Stats*) on the probability to play the optimal strategy. This probability is lower in the *Info* treatment and increases when the subject is younger, knows some basics of statistics and has gained some experience with the game.

The negative effect of information is tightly linked to the overall effect of *Info* on the number of answers (out of six) which are hidden by subjects. On average, subjects hide more answers in *Info* than in *Control*, respectively 1.97 and 1.88 answers ($p = 0.064$). The frequencies with which subjects hide different numbers of answers is given in Figure 2. The distributions of these frequencies are significantly different in *Control* and *Info* according to the Kolmogorov-Smirnov test ($p < 0.001$). Mainly, we see significantly fewer subjects hide one answer and significantly more subjects hide two or three answers in *Info* than in *Control* (all differences being significant at the 1% level). Said differently,

https://doi.org/10.1017/eec.2025.10030 Published online by Cambridge University Press

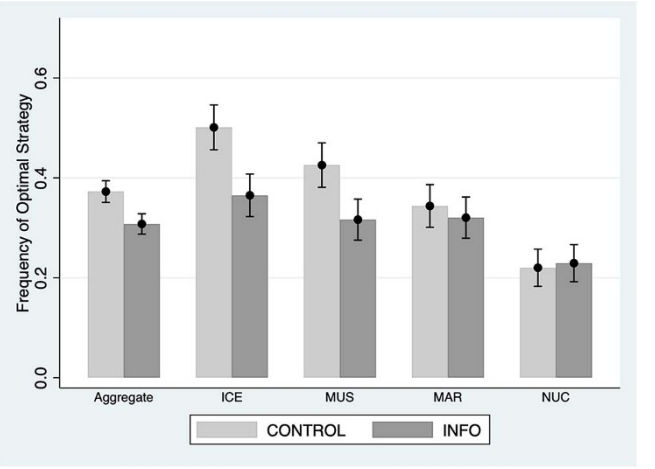

Fig. 3 Frequency of optimal strategy, per treatment and per target

the *Info* increases the share of subjects who hide two or three answers, a sub-optimal choice in 65.15% of the cases. An interpretation is that, overall, information about the functioning of the algorithm makes subjects over-think and look for more correlations than there truly are. This interpretation is reinforced by the reports subjects make in Part 3: In *Control*, subjects answer that the target question is correlated to no other question 40.46% of the time while they give this answer only 31.95% of the time in *Info*, a significant difference ($p < 0.001$).

Result 1. *Pooling all target questions, the frequency of optimal strategies is lower in Info than in Control. Fewer subjects hide one answer and more subjects hide two or three answers in Info than in Control.*

3.3. Effect of target questions

In this section, we unpack result 1 for each target and examine the validity of hypothesis 2.

Figure 3 gives the frequency of optimal strategy in *Info* and *Control* for each target question separately. Conditional on each treatment, subjects play the optimal strategy significantly less frequently when the target question is NUC than when it is any other question (all p-values are smaller than 0.002). This finding validates hypothesis 2 and suggests that subjects play more optimally when it is easier for them to identify the correlation or absence of correlation between the target question and other questions.

Result 2. *Conditional on the information subjects have about the functioning of the algorithm, the frequency of optimal strategy is lower when the target question is NUC than when the target is any other question.*

For the two uncorrelated targets, the frequency of optimal strategies is significantly higher in *Control* than in *Info*: It falls from 50.10% to 36.51% for ICE ($p < 0.001$) and from 42.56% to 31.64% for MUS ($p < 0.001$). Table 7 in section A.3 of the Appendix provides the regressions confirming this finding. This decline coincides with an increase in the average number of answers hidden by subjects: The average increases from 1.68 to 1.83 for ICE ($p = 0.089$), and from 1.78 to 1.96 for MUS ($p = 0.054$).[16] In fact, the above-stated Result 1 is importantly driven by how subjects play with

[16]Considering the two uncorrelated targets, the Kolomogorov-Smirnov test establishes that the distributions of the frequencies with which subjects hide from 0 to 6 answers are different in *Control* and *Info* ($p = 0.001$).

https://doi.org/10.1017/eec.2025.10030 Published online by Cambridge University Press

the uncorrelated targets: In *Info*, subjects see more correlations than there are, which significantly decreases the share of subjects who hide the target only. In Part 3, 53.35% of subjects properly identify that the target is correlated to no other question in *Control* while this number decreases to 43.10% in *Info* ($p < 0.001$).

Result 3. *When the target question is uncorrelated, the frequency of optimal strategies is lower in Info than in Control. In the former treatment, subjects search for nonexistent correlations and hide more answers than what is optimal.*

As it appears on Figure 3, when the target is MAR or NUC, there is no statistically significant difference in the frequency of optimal strategy between *Control* and *Info*. Table 8 in section A.3 of the Appendix provides the regressions confirming this finding. For MAR, this frequency is 34.38% in *Control* versus 32.05% in *Info* ($p = 0.441$). For NUC, this frequency is 22.01% in *Control* versus 22.92% in *Info* ($p = 0.735$). Pooling MAR and NUC, the average number of hidden answers are not different in *Control* and *Info*, respectively 2.03 and 2.04 answers ($p = 0.938$).[17] However, it is hard to interpret the absence of treatment effect for correlated questions because it hides very important difference between the MAR and NUC target questions, and between the way common and uncommon subjects play with these questions. We describe these differences in detail in the next section.

3.4. Effect of being a common or an uncommon subject

In this section, we examine how common and uncommon subjects play when the target questions are MAR and NUC.[18] We remind the reader that the optimal strategy of common subjects is to hide their answers to the target question and to the most correlated question whereas the optimal strategy of uncommon subjects is to hide only their answer to the target (Proposition 2). We will see that, when the target is MAR, common and uncommon subjects reach similar frequencies of optimal strategies, necessarily by making different disclosure choices. In contrast, when the target is NUC, common and uncommon subjects make similar disclosure choices, thereby reaching different frequencies of optimal strategies.

One important reason behind these findings is that MAR and NUC are very different correlated target questions. On the one hand, 79.69% of subjects (pooling *Control* and *Info*) correctly identify that the question about being married is correlated to the question about having children. In addition, it is very likely that, by identifying this correlation, subjects also directly identify its direction: Being married is correlated to having children, not to having no children. On the other hand, only 26.39% of subjects (pooling *Control* and *Info*) correctly identify that the question about the use of nuclear power is correlated to gender. And, if subjects identify this correlation correctly, its direction may not be obvious.

3.4.1. The MAR target question

When the target is MAR, common and uncommon subjects play differently. This is shown on Figures 4(a) and (b) which display, for common and uncommon subjects separately, the frequencies with which they hide 0 to 6 answers in each treatment.

We start with the *Control* treatment. In this treatment, common and uncommon subjects reach similar share of optimal strategies (34.35% and 34.48% respectively, $p = 0.979$), which they do by making different disclosure choices. This is visible by looking at the light gray bars on both sides

[17] Considering the two correlated targets, the Kolomogorov-Smirnov test establishes that the distribution of the frequencies with which subjects hide from 0 to 6 answers are not different in *Control* and *Info* ($p = 0.225$).

[18] This part of the analysis is exploratory as we did not pre-register any hypothesis about how these two types of subjects would play.

https://doi.org/10.1017/eec.2025.10030 Published online by Cambridge University Press

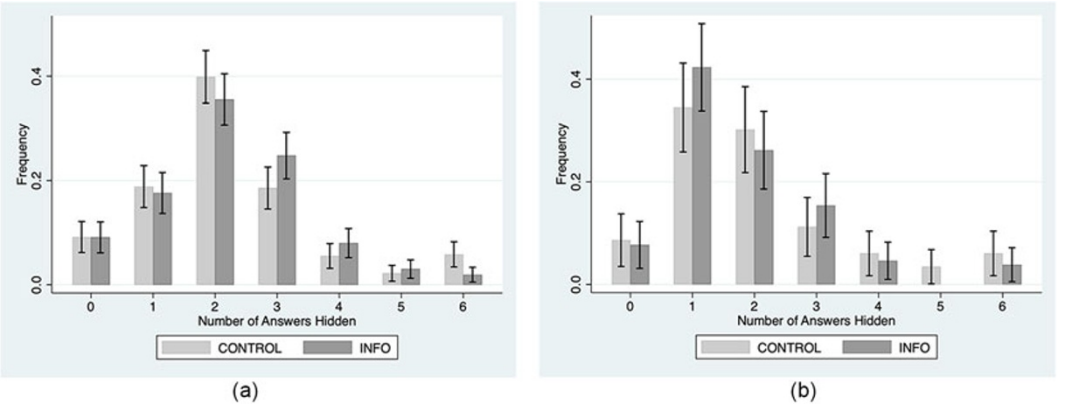

Fig. 4 Hiding 0 to 6 answers for the MAR target, per treatment. (a) Common subjects. (b) Uncommon subjects

of Figure 4: Only 18.84% of common subjects hide one answer against 34.48% of uncommon subjects ($p < 0.001$); 39.89% of common subjects hide two answers against 30.17% of uncommon subjects ($p = 0.060$).[19] Clearly, when uncommon subjects hide the target question only (the optimal strategy for them), it could be either because they use this relatively natural strategy without thinking much or because they are sophisticated enough to do so to mislead the algorithm. These two possibilities are confounded in our data. The number of common subjects who hide only the target (a sub-optimal strategy for them) is 14.96%. If we consider this number as a benchmark for the fraction of subjects who use this strategy simply because it is natural, it leaves about 20% of uncommon subjects (a significant difference between 34.48% and 14.96%, $p < 0.001$) who use this strategy because they understand that disclosing their answer to CHI misleads the algorithm. Overall, the data in *Control* suggests that, when subjects have well understood a correlation, a significant share of them is sophisticated enough to strategically play with it against the algorithm.

Next, we consider the *Info* treatment. For common subjects, as well as for uncommon subjects, there is no significant effect of the *Control* versus *Info* treatment on the frequency of optimal strategies or on the number of hidden answers. For common subjects, the share of optimal strategy is 34.35% in *Control* and 29.48% in *Info* ($p = 0.160$), and they hide an average of 2.22 answers in both treatments ($p = 0.947$). For uncommon subjects, the share of optimal strategy is 34.48% in *Control* and 39.23% in *Info* ($p = 0.443$), and they hide an average of about two answers in both treatments ($p = 0.188$). Our interpretation is that, when the correlation is obvious and identified by most subjects, it does not bring much to subjects to learn that the algorithm uses correlations.

Second, the small, insignificant effect of *Info* on subjects' disclosure strategies goes in different directions for common and uncommon subjects. It follows, as shown on Figure 4, that the difference in play between common and uncommon subjects is even larger for *Info* than for *Control*.[20] In *Info*, the share of optimal strategies for common subjects is significantly lower than the share for uncommon subjects (29.48% against 39.23%, $p = 0.041$). In fact, the *Info* treatment pushes common subjects in the same direction as the one identified earlier for uncorrelated targets: They start thinking about nonexistent correlations and hide more than what is optimal. In particular, 24.79% of common subjects sub-optimally hide three answers in *Info* against 18.56% in *Control* ($p = 0.042$). For uncommon subjects, the *Info* treatment pushes subjects in the other direction in that more subjects hide one

[19]For *Control*, the Kolomogorov-Smirnov test confirms that the distributions of frequencies with which common and uncommon subjects hide 0 to 6 answers are different ($p = 0.036$).

[20]For *Info*, the Kolomogorov-Smirnov test confirms that the distributions of frequencies with which common and uncommon subjects hide 0 to 6 answers are different ($p < 0.001$).

https://doi.org/10.1017/eec.2025.10030 Published online by Cambridge University Press

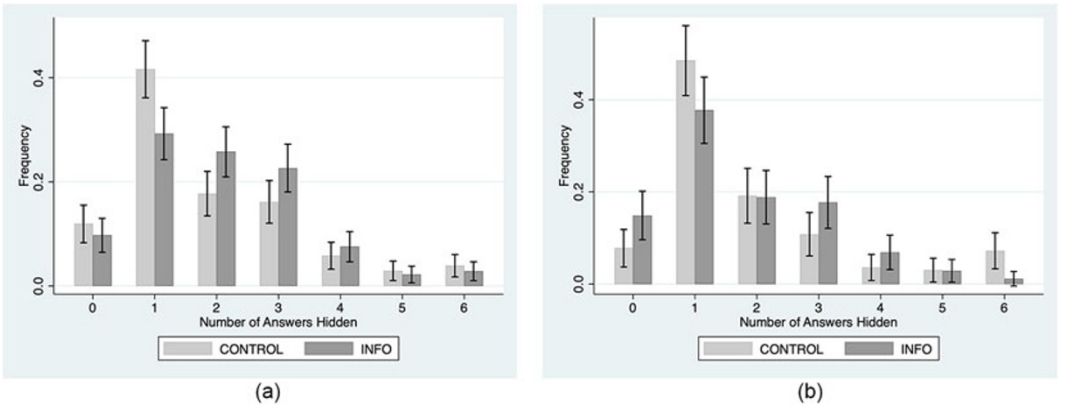

Fig. 5 Hiding 0 to 6 answers for the NUC target, per treatment. (a) Common subjects. (b) Uncommon subjects

answer only (42.31% in *Info* versus 34.48% in *Control*, $p = 0.210$). This suggests that the *Info* treatment not only pushed to think about correlations but also pushes some subjects to think about the direction of these correlations, and to play slightly better as uncommon subjects.

Result 4. *When the target question is correlated and the correlation is well-understood by most subjects, common subjects play differently from uncommon subjects, both in Control and in Info. For each group of subjects, there is no significant effect of the Control versus Info treatment on the frequency of optimal strategy.*

3.4.2. The NUC target question

When the target is NUC, common and uncommon subjects play similarly. This appears on Figures 5(a) and (b) which display, for common and uncommon subjects separately, the frequencies with which they hide 0 to 6 answers in each treatment.

We start with the *Control* treatment. Looking at Figure 5, we see that common and uncommon subjects play similarly, most often hiding one answer only (41.61% of common subjects and 48.50% of uncommon subjects do so, $p = 0.149$).[21] This is linked to the fact that 46.71% of common subjects and 43.23% of uncommon subjects answer that NUC is correlated to no other question in Part 3 of the experiment. The similar disclosure strategies used by common and uncommon subjects result in very different shares of optimal strategies: 8.06% of common subjects play optimally against 47.90% of uncommon subjects ($p < 0.001$).

Next, we find that the *Info* treatment importantly affects the beliefs about correlations reported in Part 3 of the experiment. In *Control*, only 20.96% of subjects correctly report that GEN is correlated to NUC. This share goes to 31.64% in the *Info* treatment ($p < 0.001$). In parallel, the share of subjects who answer that NUC is not correlated to any other question decreases from 44.44% in *Control* to 32.45% in *Info* ($p < 0.001$). As it appears on Figure 5, these changes in beliefs go with a decrease in the share of subjects (common and uncommon) who hide one answer only. This share goes from 44.03% in *Control* to 32.25% in *Info* ($p < 0.001$). We also observe a significant increase in the share of subjects who hide three answers ($p = 0.007$). These changes are importantly driven by a higher fraction of subjects hiding their gender in *Info* (41.38% against 30.19% in *Control*, $p < 0.001$). Since common and uncommon subjects react similarly to *Info* by hiding more answers, the frequency of optimal strategies increases for common subjects and decreases for uncommon subjects. This is summarized

[21] For *Control*, the Kolomogorov-Smirnov test confirms that the distributions of frequencies with which common and uncommon subjects hide 0 to 6 answers are not different ($p = 0.992$).

https://doi.org/10.1017/eec.2025.10030 Published online by Cambridge University Press

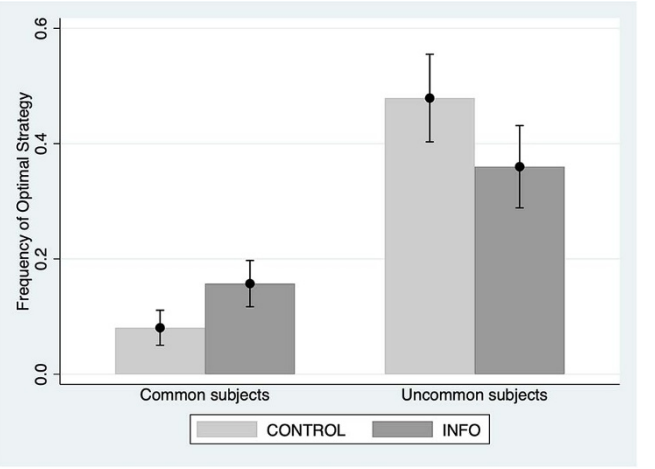

Fig. 6 Frequency of optimal strategies for NUC, per treatment and type of subject

on Figure 6. Overall, this suggests that, while the *Info* treatment helps subjects identify better which question is correlated to the target, it does not necessarily help them to understand the direction of this correlation and to play better as uncommon subjects.

Result 5. *When the target question is correlated but the correlation is hard to identify, common and uncommon subjects play similarly. Subjects identify better the correlation in Info than in Control. The frequency of optimal strategies is higher in Info than in Control for common subjects, and lower in Info than in Control for uncommon subjects.*

3.5. Consistent strategies

In previous sections, we examine subjects' behavior by considering the optimality of their disclosure strategies. In this section, we take another approach and evaluate the consistency between subjects' disclosure strategies and the correlations they report in Part 3 of the experiment. Because we try to get at the logic behind disclosure choices, we first restrict our dataset to the strategies for which subjects hide the answer to the target question. We are left with 3102 observations (for this subsection). Next, we define a *consistent strategiy*:

- When a subject reports no correlation between the target question and any of the other questions (in Part 3), we say that his/her strategy is *consistent* if he/she hides only the target question; the strategy is *inconsistent* otherwise.
- When a subject reports a correlation between the target question and another question (in Part 3), we say that his/her strategy is *consistent* if he/she hides the answer to that other question and the answer to the target; the strategy is *inconsistent* otherwise.

A number of remarks about the definition of consistent strategies are in order. First, the definition is independent of whether or not the reports made in Part 3 are correct. Second, we consider that a subject who, for example, reports that NUC is most correlated to GEN uses a consistent strategy if he/she hides the answer to NUC and GEN independently of whether he/she hides other answers. Recall that, in Part 3, subjects can only report the question they consider is most correlated to the target, not all questions they believe are correlated. Third, note that the definition of consistency does not consider the distinction between common and uncommon subjects. We define consistency as

https://doi.org/10.1017/eec.2025.10030 Published online by Cambridge University Press

Table 3. Frequency of consistent strategies, by treatment and target

	Control	*Info*	Overall	p-value
ICE	57.40	52.21	54.81	0.148
MUS	52.27	50.00	51.11	0.531
MAR	62.47	61.10	61.77	0.692
NUC	52.32	52.96	52.64	0.859
Total	56.15	54.12	55.13	0.257

Table 4. Frequency of correctly identified correlations, by treatment and target

	Control	*Info*	Overall	p-value
ICE	52.99	45.45	49.22	0.037
MUS	52.00	37.44	44.58	$<$0.001
MAR	80.46	83.54	82.03	0.261
NUC	22.16	33.42	27.80	$<$0.001
Total	51.92	50.22	51.06	0.345

the act of hiding the answer to a question identified as most correlated to the target, ignoring that it is sub-optimal for uncommon subjects. Said differently, we consider as inconsistent the optimal strategies of uncommon subjects.

Table 3 gives the frequency of consistent strategies per treatment and target question. Pooling all treatments and targets, 55.13% of strategies are consistent. In addition, for a given target, the level of consistency is never affected by the *Control versus Info* treatment.

Overall, strategies are more consistent (55.13%) than they are optimal (33.97%). Playing optimal strategies requires that subjects correctly understand, among other things, the correlations between the targets and other questions. In Table 4, we give the frequency of correct answers given in Part 3, that is, the share of cases in which subjects see that CHI is correlated to MAR, that GEN is correlated to NUC, and that no question is correlated to ICE and MUS (see Appendix A.2 for details for the whole sample). For uncorrelated targets, *Info* makes subjects less accurate. For correlated targets, *Info* has no effect on the reported correlation for MAR but helps subjects identify the correct correlation for NUC.

Combining a relatively constant level of consistency with reports about correlations that vary across treatments, we get the following result. It can partly explain why the overall effect of information on the optimality of disclosure strategies is negative.

Result 6. *More than half of the disclosure strategies are consistent with the correlations that subjects report, and the level of consistency is independent of the treatment. The treatment however affects the accuracy of reported correlations: When the target is uncorrelated, subjects act on less accurate reports in Info than in Control; when the target is correlated, subjects act on similar or more accurate reports in Info than in Control.*

4. Conclusion

We propose an experiment in which subjects strategically disclose multi-dimensional information about themselves to a Naive Bayes algorithm trained to deduce non-disclosed attributes from disclosed ones. In an experimental variation, we explain to the subjects that the algorithm uses existing correlations between attributes to make deductions. Such information about the functioning of the algorithm affects subjects' behavior in ways that importantly depend on what they initially know

https://doi.org/10.1017/eec.2025.10030 Published online by Cambridge University Press

about existing correlations:When subjects rightly expect no correlations between some attributes, the information makes them overthink and disclose less optimally; when correlations are obvious, the information has little effect on disclosure strategies; when correlations are hard to see, information helps subjects identify these correlations but not necessarily their directions.

A central message of our work is that, in order to play optimally against algorithms trained on large datasets, individuals need more than transparency about the functioning of the algorithms — they must understand the underlying statistical structure of the data on which the algorithms rely. A large body of research has documented behavioral biases, such as base-rate neglect (Kahneman and Tversky, 1973) and correlation neglect (Enke and Zimmermann, 2019), as well as cognitive limitations like limited memory (Wilson, 2014). These factors can influence individuals' beliefs about the characteristics present in the training data and the correlations between them. Recent literature in behavioral economics explores how people form mental models of variable relationships based on personal observations (Fréchette et al., 2025). In our experiment, we inform participants that the algorithm relies on correlations, which encourages them to think about these relationships. However, we do not explicitly design a treatment to prompt them to consider the nature of the training data.

Our work further shows that individuals additionally need to understand how the characteristics of the training dataset relate to their own characteristics. In the well-structured setting we consider, it is possible to characterize optimal disclosure strategies for all subjects, that is, for all their possible sets of characteristics. This characterization demonstrates that the distinction between common and uncommon subjects is crucial: Subjects whose characteristics are not mainstream can trick the algorithm into making wrong guesses about their characteristics precisely because the algorithm is trained on large data. This observation raises novel questions about how subjects perceive themselves in relation to others. In the experimental psychology literature, Ross et al. (1977) shows that individuals are biased towards seeing more consensus about their (hypothetical) decisions or characteristics than there is. Our study suggests that it is important to understand the extent to which people can identify the traits that make them different from or similar to the crowd.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/eec.2025.10030.

Replication package. The replication material for the study is available at https://zenodo.org/records/17085920.

Acknowledgements. We thank Victor Augias, Emeric Henry, Frédéric Koessler, Theo Marquis, Eduardo Perez-Richet, Franz Ostrizek and seminar participants at Sciences Po and CREST for helpful suggestions. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n°850996 – MOREV). The study was pre-registered in 2023 with AsPredicted (#128706) and received ethical approval from Sciences Po, France.

References

Acquisti, A. (2004). Privacy in electronic commerce and the economics of immediate gratification. Proceedings of the 5th ACM conference on Electronic commerce, New York, NY, USA, 21–29.

Acquisti, A., Adjerid, I., Balebako, R., Brandimarte, L., Cranor, L. F., Komanduri, S., Leon, P. G., Sadeh, N., Schaub, F., Sleeper, M. et al. (2017). Nudges for privacy and security: Understanding and assisting users' choices online. *ACM Computing Surveys (CSUR)*, *50*(3), 1–41.

Acquisti, A., Brandimarte, L., & Loewenstein, G. (2020). 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.

Acquisti, A., John, L., & Loewenstein, G. (2012). The impact of relative standards on the propensity to disclose. *Journal of Marketing Research*, *49*(2), 160–174.

Ball, I. (2024). Scoring strategic agents. *American Economic Journal: Microeconomics*, *17*(1), 97–129. February 2025.

Basu, C., Hirsh, H., & Cohen, W. (1998). Recommendation as classification: using social and content-based information is recommendation. Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence, Madison, Wisconsin, USA, 714–720.

Björkegren, D., Blumenstock, J. E., & Knight, S. (2020). Manipulation-proof machine learning. (arXiv:2004.03865).

Bó, I, Chen, L., & Hakimov, R. (2023). Strategic responses to personalized pricing and demand for privacy: An experiment. *Working Paper.*

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825.

Dargnies, M. P., Hakimov, R., & Kübler, D. (2024). Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence. *Management Science*.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170.

Eliaz, K., & Spiegler, R. (2019). The model selection curse. *American Economic Review: Insights*, *1*(2), 127–140.

Eliaz, K., & Spiegler, R. (2022). On incentive-compatible estimators. *Games and Economic Behavior*, *1*(2), 204–220.

Enke, B., & Zimmermann, F. (2019). Correlation neglect in belief formation. *The Review of Economic Studies*, *86*(1), 313–332.

Frankel, A., & Kartik, N. (2022). Improving information from manipulable data. *Journal of the European Economic Association*, *20*(1), 79–115.

Fréchette, G. R., Vespa, E., & Yuksel, S. (2025). Extracting statistical relationships from observational data: Predicting with full or partial information. *AEA Papers and Proceedings*, *115*, 637–642.

Hardt, M., Megiddo, N., Papadimitriou, C., & Wootters, M. (2016). Strategic classification. Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science, Cambridge, Massachusetts, USA, 111–122.

Hu, L., Immorlica, N., & Vaughan, J. W. (2019). The disparate effects of strategic manipulation. Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, Georgia, USA, 259–268.

John, L., Acquisti, A., & Loewenstein, G. (2011). Strangers on a plane: Context-dependent willingness to divulge sensitive information. *Journal of Consumer Research*, *37*(5), 858–873.

Jussupow, E., Benbasat, I, & Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. *Publications of Darmstadt Technical University, Institute for Business Studies*.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237.

Kamel, H., Abdulah, D., & Al-Tuwaijari, J. M. (2019). Cancer classification using gaussian naive bayes algorithm. 2019 International Engineering Conference (IEC), Erbil, Iraq, 165–170.

Kennedy, B., Funk, C., & Tyson, A. (2023). Majorities of americans prioritize renewable energy, back steps to address climate change: But many foresee problems ahead with transition to renewables and oppose breaking from fossil fuels altogether. Technical report Pew Research Center.

Kleinberg, J., & Raghavan, M. (2020). How do classifiers induce agents to invest effort strategically?. *ACM Transactions on Economics and Computation*, *8*(4), 1–19.

Krishnaswamy, A., Li, H., Rein, D., Zhang, H., & Conitzer, V. (2021). Classification with strategically withheld data. *Proceedings of the AAAI Conference on Artificial Intelligence, Virtual conference*, *35*(6), 5514–5522.

Meir, R., Procaccia, A., & Rosenschein, J. (2012). Algorithms for strategyproof classification. *Journal of Artificial Intelligence*, *186*, 123–156.

Miklós-Thal, J., Goldfarb, A., Haviv, A., & Tucker, C. (2024). Frontiers: Digital hermits. *Marketing Science*, *43*(4), 697–708.

Pal, R., Shaikh, S., Satpute, S., & Bhagwat, S. (2022). Resume classification using various machine learning algorithms. *ITM Web Conf.*, *44*, 03011.

Perez-Richet, E., & Skreta, V. (2022). Test design under falsification. *Econometrica*, *90*(3), 1109–1142.

Pronk, V., Verhaegh, W., Proidl, A., & Tiemann, M. (2007). Incorporating user control into recommender systems based on naive bayesian classification. Proceedings of the 2007 ACM conference on Recommender systems, Minneapolis, Minnesota, USA, 73–80.

Rawat, S., Rawat, A., Kumar, D., & Sabitha, A. S. (2021). Application of machine learning and data visualization techniques for decision support in the insurance sector. *International Journal of Information Management Data Insights*, *1*(2), 100012.

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, *13*(3), 279–301.

Sahu, S., Nautiyal, A., & Prasad, M. (2017). Machine learning algorithms for recommender system - a comparative analysis. *International Journal of Computer Applications Technology and Research*, *6*(2), 97–100.

Slokom, M., Hanjalic, A., & Larson, M. (2021). Towards user-oriented privacy for recommender system data: A personalization-based approach to gender obfuscation for user profiles. *Information Processing & Management*, *58*(6), 102722.

Solomon, L. S., Tomaskovic-Devey, D., & Risman, B. J. (1989). The gender gap and nuclear power: Attitudes in a politicized environment. *Sex Roles*, *21*(5), 401–414.

Valdiviezo-Diaz, P., Ortega, F., Cobos, E., & Lara-Cabrera, R. (2019). A collaborative filtering approach based on naïve bayes classifier. *IEEE Access*, *7*, 108581–108592.

Wang, K., & Tan, Y. (2011). A new collaborative filtering recommendation approach based in naive bayes classifier. Proceedings of the Second international conference on Advances in swarm intelligence - Volume Part II, Chongqing, China, 218–227.

https://doi.org/10.1017/eec.2025.10030 Published online by Cambridge University Press

Wilson, A. (2014). Bounded memory and biases in information processing. *Econometrica*, *82*(6), 2257–2294.
Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z. -H., Steinbach, M., Hand, D. J., & Steinberg, D. (2007). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1–37.

Cite this article: Hagenbach, J., & Salas, A. (2025). Strategic information disclosure to classification algorithms: an experiment. *Experimental Economics*, 1–22. https://doi.org/10.1017/eec.2025.10030