Strategic Information Disclosure to Recommendation Algorithms: An Experiment*

Jeanne Hagenbach† Aurélien Salas‡

September 26, 2024

Abstract

We experimentally study individuals' ability to 'game' recommendation systems. Subjects initially answer a few binary questions about themselves. Next, they play several rounds of a game during which they must decide whether to disclose or hide their answers to a Naive Bayes algorithm programmed to guess these answers. In each round, their objective is to minimize the algorithm's accuracy in guessing the answer to a specific question, the target. We exogenously vary the information provided to participants about the functioning of the algorithm, and how easy it is for subjects to identify the correlations that exist between the target question and other questions. Over all rounds, subjects play the optimal disclosure strategy less frequently when informed that the algorithm uses correlations to guess answers. This information makes subjects overthink about correlations; they identify correlations they otherwise do not see but also see non-existent ones. Information does not significantly help subjects realize the directions of existing correlations, which turn out to be key to playing optimally against the algorithm.

Keywords: Experiments, strategic disclosure, online information, recommendation systems.
JEL classification: C91, D89, D91, M38

*We thank Victor Augias, Emeric Henry, Frédéric Koessler, Theo Marquis, 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).

†CNRS, Sciences Po, WZB, CEPR, CESifo - jeanne.hagenbach@sciencespo.fr

‡Sciences Po - aurelien.salas@sciencespo.fr

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 can 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 an algorithm trained on other individuals' data to guess their personal attributes, a technique - collaborative filtering - at the core of most recommendation systems. The setting we consider is particularly simple for subjects compared to real-life situations: their objective is clearly defined - they must prevent the algorithm from guessing one specific attribute - and their task consists in strategically disclosing or 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.[1] 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.[2] In the second part, subjects play against an algorithm which does not know their individual characteristics but has been trained to guess them. More precisely, Part 2 is made up of four, independent rounds of

[1]The study was pre-registered in 2023 with AsPredicted, #128706.

[2]In sectionA.4 of the Appendix, we explain in detail how we selected this particular set of questions.

play. In each round, 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 (there is no possibility to lie). Experimental payments in this game represent a trade-off commonly faced by individuals in front of recommendation systems. On the one hand, hiding characteristics is costly - it takes time and effort to prevent algorithms from accessing personal data -, so we reduce players' initial endowment for every hidden answer. On the other hand, hiding characteristics reduces the information that the algorithm can use to guess subjects' characteristics. In the game we propose, the subjects' payment is inversely proportional to the accuracy with which the algorithm guesses their answer to the target question. In the third part of the experiment, we ask subjects to report the correlations they see between the target questions and the other questions of Part 1.

The algorithm we use in the experiment 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, the 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 in which they answered the same questions as our main subjects. Their answers serve as training data for the algorithm. In short, the algorithm of our experiment 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, each subject plays four rounds, with four different target questions. This allows to consider both strong and absent correlations between target questions and other questions, as well as vary how obvious the correlations are when they exist. Precisely, we consider the following four target questions: two questions, abbreviated ICE and MUS for favorite ice-cream flavor and time spent listening to music, whose answers are not correlated to any other answers given by subjects in Part 1; the question about being married, 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 the gender. This last correlation is harder to identify for subjects, as indicated by the data gathered in Part 3 of the experiment.

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, characterized for every subject and target question, turn out to be relatively intuitive. Intuitively, their characterization comes down to finding the largest set of characteristics which can be disclosed without increasing the accuracy of the algorithm guess by too much. 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 *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 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 for them to game the algorithm.[3] To play optimally, subjects need to understand the functioning of the algorithm, existing correlations but also the directions of these correlations.

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 information 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 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. With this target question, subjects appear sophisticated

[3]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.

enough to play differently when being common or uncommon, that is, to strategically disclose or hide whether they have children. Finally, 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 common vs. uncommon distinction, for which they additionally need to understand the direction of this correlation, namely that it is being a male (not a female) which is correlated to being in favor of nuclear power. Subjects hide their gender more frequently in *Info* than in *Control*, but they do so independently of whether it is optimal (because they are common) or sub-optimal (because they are uncommon).

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.[4] We study individuals' strategies for a fixed algorithm rather than adapt the algorithm to these strategies.

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 induce more truthful revelation from agents who want higher allocations or scores. In a different type of work, Eliaz and Spiegler (2019, 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

[4]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.

penalties for including variables in the model can create incentives for the agent to misreport his characteristics. In all the above-mentioned 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 disclosure 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), etc. 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.

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 United States) who completed a very simple task: they answered 30 binary questions about themselves.[5] Subjects were paid a fixed amount of £0.6 for filling out that questionnaire, which took them 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 can earn money in each part as detailed below.[6]

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.

[5] Only the question about gender was not binary. The pre-study questionnaire is given in section 2 of the Online Appendix.

[6] The complete instructions are given in section 2 of the Online Appendix.

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

GEN - What gender are you currently? *Male / Female / Non-Binary.*[7]

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*

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*. A round of game has 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

[7]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.

subject was "No". The probability that is computed for the true answer of the subject is called the *guess* of the algorithm.

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.

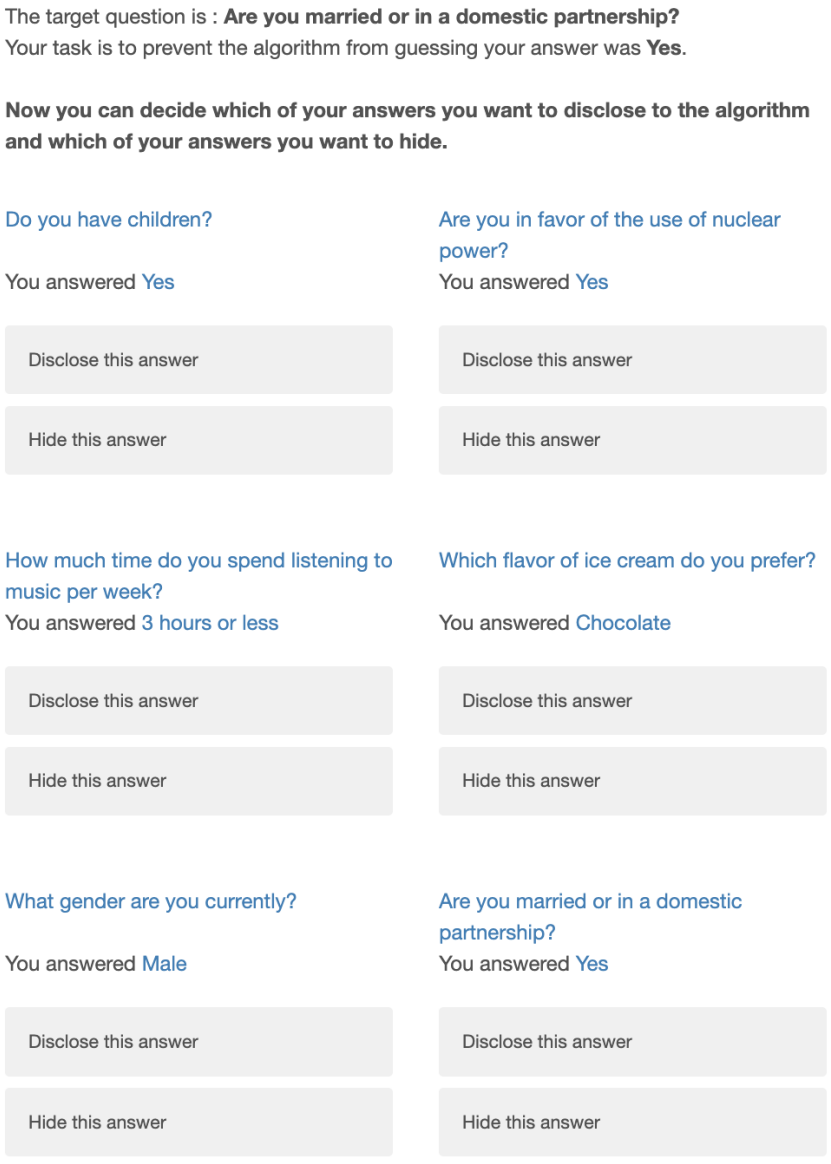


Figure 1: A screen seen by subjects when they had to make their disclosure choices

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 6 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:[8] when the target question is j and the subject discloses 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. This algorithm is widely used in practice and the assumption of conditional independence simplifies a lot the relationship between disclosed answers and the guess. Mainly, each disclosed answer contributes independently to the guess and subjects do not need to think about the effect of hiding answers, which we discuss in section A.6 of the Appendix. 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 in that dataset.

[8] Practically, we use the Bernouilli Naive Bayes code in the sklearn package available in Python. For details about how it is adapted to our setting, see section 3 of the Online Appendix.

In the particular case in which the subject discloses the answer to the target question j itself, then (1) leads to $g_D^j = 1$[9] 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.

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 £0.2. (2) At the end of the round, the endowment is reduced by two times the guess (between 0 and 1) computed by the algorithm. This way, hiding answers is costly but reduces the information available to the algorithm to make its guess. 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.

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.

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

[9]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.

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 £0.10. 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 recommendations systems, algorithms, Internet and statistics.

Implementation. The experiment was run on Prolific and involved 970 subjects (fluent in English and based in the United States).[10] 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 £0.5). The experiment was pre-registered (reference #128706 on Aspredicted) and received ethical approval from Sciences Po, France.[11]

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.

[10] Out of 982 in total, we had to drop the 12 subjects who answered *Non-Binary* to GEN (see footnote 7).

[11] We had pre-registered a third treatment. It is presented and briefly analyzed in section A.7 of the Appendix.

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.*
- 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 3 hours of music per week.*

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. Section A.4 of the Appendix gives details about the selection.

We use ICE, MUS, MAR and NUC as target questions. 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).[12] Again, the

[12]This result is not specific to our sample. Solomon et al. (1989) show that gender differences in safety concerns explain the gender gap in acceptance of nuclear power. Recently, the study by Kennedy et al. (2023) finds that "men continue to be far more likely than women" to favor the use of nuclear power.

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

2.4 Optimal Strategies

In this subsection, we derive the subjects' optimal disclosure strategies in the game they play against the algorithm. Intuitively, the exercise consists in finding, for every subject, a disclosure set which is large (hiding is costly) but 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 general results about optimal strategies for all subjects, we need to consider all possible sets of answers A that these subjects could have given. 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. Proofs of all the 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. To establish this result, we start from a set D of disclosed answers which contains the answer to the target question. Next we show that switching to hiding this answer costs £0.20 but reduces the guess from probability 1 to at least probability 0.825, across all possible sets of characteristics A, target questions and sets D. This makes a net benefit of at least £0.15.

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

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 £0.20.

Proposition 2 *In the game against the algorithm, when the target question is uncorrelated (ICE or MUS), it is optimal for every subject to hide only the answer to the target question.*

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, respectively CHI and GEN. For subjects with similar answers to the majority of the subjects in the pre-study data, disclosing these answers helps the algorithm make a better guess about their answer to the target question. For such subjects and even if hiding is costly, it is therefore beneficial to hide the answer to the question correlated to the target question. For subjects with different answers to the majority of subjects in the pre-study data, their answers mislead the algorithm about their answer to the target question. For such subjects, it is therefore beneficial to disclose the answer to the question correlated to the target question.

For each target question, we define two populations of subjects, *common* and *uncommon* subjects, depending on whether or not these subjects answered like the majority of subjects in the pre study. Note that, for a subject to identify whether he/she is common or uncommon or, equivalently, to determine whether the answer to the correlated question should be disclosed or hidden, he/she needs to understand the directions of existing correlations: it is having

children (not having no children) which is most correlated to being married; it is being a male (not a female) which is most correlated to being in favor of nuclear power.

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 the two types of subjects.

Proposition 3 *In the game against the algorithm, when the target question is correlated (MAR or NUC), it is optimal for every common subject to hide exactly two answers: the answer to the target question and the answer to its correlated question (resp. CHI or GEN).*

Proposition 4 *In the game against the algorithm, 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.*

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 last line gives, for each target, the percentage of cases in which the optimal disclosure strategy is to hide only the answer to the target question. Note that, for MAR and NUC, this percentage corresponds to the fraction of uncommon subjects.

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

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 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. 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 36 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.[13] 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 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.

[13] Subjects' effective payoffs in Part 2 of the experiment cannot be used directly to study subjects' behavior. These payoffs importantly depend on the frequencies of the different characteristics (and conditional characteristics) in the pre-study data. Think, for instance, of the target question being whether or not a subject listens to more than 3 hours of music weekly. Two subjects who answered differently but both disclose nothing to the algorithm do not reach the same payoffs simply because the priors used by the algorithm are not uniform: in the pre-study, 60% of subjects listen to 3h or more and about 40% do not. Looking at the difference between subjects' payoffs and the maximal payoffs they could obtain given the training data, we find the same qualitative results as in the main analysis.

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

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 number 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, 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 ($p < 0.001$).

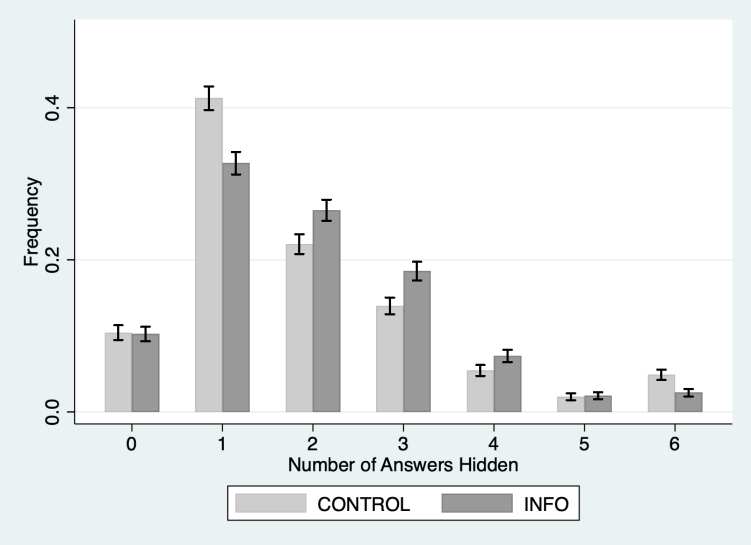


Figure 2: Hiding 0 to 6 answers, per treatment and pooling all targets

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 6 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.*

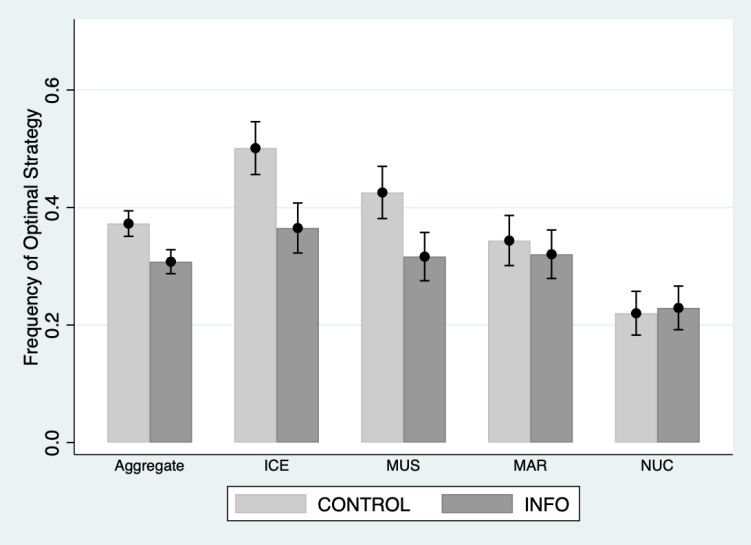


Figure 3: Frequency of optimal strategy, per treatment and per target

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 ($p < 0.001$).[14] 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 ($p = 0.054$).[15] In fact, the above-stated Result 1 is importantly driven by how subjects play with 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*.[16] For

[14] Table 5 in section A.3 of the Appendix provides the regressions confirming this finding.

[15] 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* ($p = 0.001$).

[16] Table 6 in section A.3 of the Appendix provides the regressions confirming this finding.

MAR, this frequency is 34.38% in *Control* vs. 32.05% in *Info* ($p = 0.441$). For NUC, this frequency is 22.01% in *Control* vs. 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 (Proposition 3) whereas the optimal strategy of uncommon subjects is to hide only their answer to the target (Proposition 4). 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.

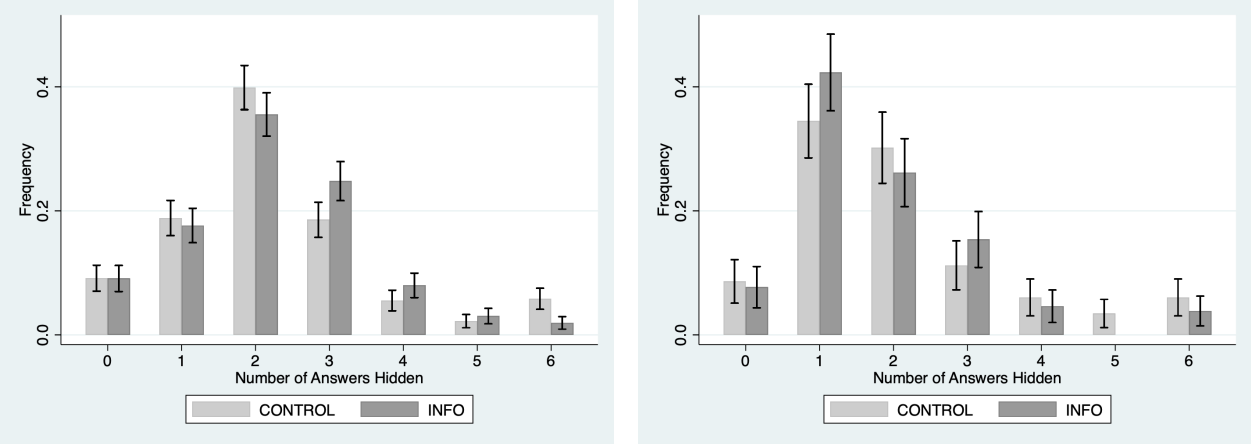
[17] Considering the two correlated targets, the Kolomogorov-Smirnov test establishes that the distribution of the frequencies with which subjects hide from 0 to 6 answers are not different in *Control* and *Info* ($p = 0.225$).

[18] We did not pre-register any hypothesis about how these two types of subjects would play because we did not foresee the impact of these types.

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.



(a) Common subjects (b) Uncommon subjects

Figure 4: Hiding 0 to 6 answers, per treatment for the MAR target

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 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

[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 ($p = 0.036$).

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* vs. *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 answer only (42.31% in *Info* versus 34.48% in *Control*, $p = 0.210$). This suggests that the *Info* treatment

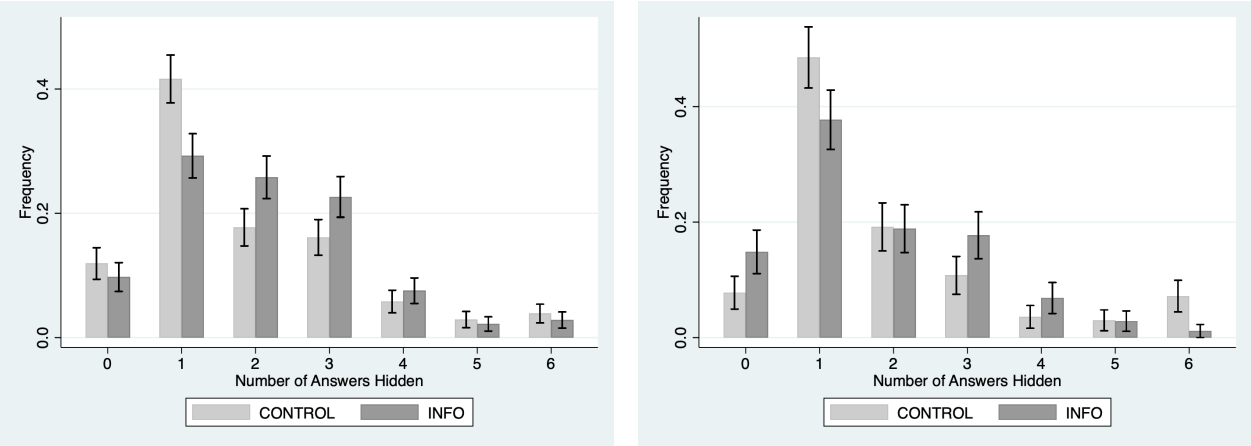
[20]For *Info*, the Kolomogorov-Smirnov test confirms that the distributions of frequencies with which common and uncommon subjects hide 0 to 6 answers are different ($p < 0.001$).

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 the Control and in the Info treatment. For each group of subjects, there is no significant effect of the Control vs. 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.



(a) Common subjects (b) Uncommon subjects

Figure 5: Hiding 0 to 6 answers, per treatment for the NUC target

We start with the *Control* treatment. Looking at the left and right parts of 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

[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 ($p = 0.992$).

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 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.

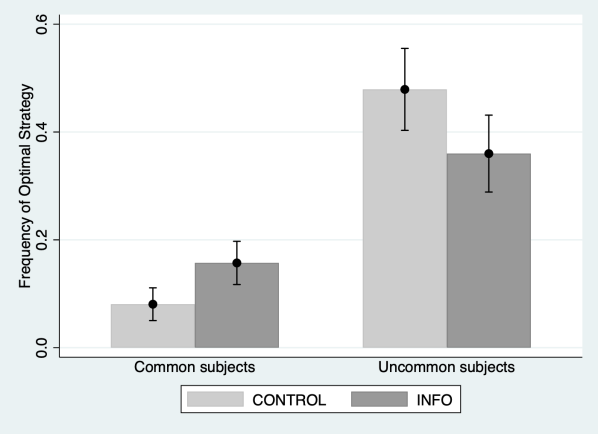


Figure 6: Frequency of optimal strategies, per treatment and type of subject

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 the Info treatment than in the Control treatment. This results in a higher frequency of optimal strategies in Info than in Control for common subjects, and in a lower frequency in Info than in Control for uncommon subjects.*

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 subjects that the algorithm uses existing correlations between attributes to make deductions. Such information about the functioning of the algorithm affects subjects' behavior in a way that importantly depends on what they initially know about existing correlations: when subjects rightly expect no correlations between some attributes, the information makes them overthink and disclose less optimally; when correlations between some attributes are obvious, the information has little effect on disclosure strategies; when correlations between some attributes are hard to see, information helps subjects identify these correlations but not necessarily their directions.

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 important: subjects whose characteristics are not mainstream can trick the algorithm into making wrong guesses about their characteristics precisely because the algorithm is trained on mainstream data. This observation raises novel questions about how subjects perceive themselves in relation to others. In particular, it is not clear to which extent people can identify the traits that make them different from the crowd, or the traits which commonly go together.

References

Acquisti, A. (2004). Privacy in electronic commerce and the economics of immediate gratification. In *Proceedings of the 5th ACM conference on Electronic commerce*, pages 21–29. Association for Computing Machinery.

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., and Loewenstein, G. (2020). Secrets and Likes: The Drive for Privacy and the Difficulty of Achieving It in the Digital Age. *Journal of Consumer Psychology*, 30:736–758.

Acquisti, A., John, L., and Loewenstein, G. (2012). The Impact of Relative Standards on the Propensity to Disclose. *Journal of Marketing Research*, 49:160–174.

Ball, I. (2024). Scoring Strategic Agents. *American Economic Journal: Microeconomics*, Forthcoming.

Bó, I., Chen, L., and Hakimov, R. (2023). Strategic Responses to Personalized Pricing and Demand for Privacy: An Experiment. *Working Paper*.

Eliaz, K. and Spiegler, R. (2019). The Model Selection Curse. *American Economic Review: Insights*, 1(2):127–140.

Eliaz, K. and Spiegler, R. (2022). On incentive-compatible estimators. *Games and Economic Behavior*, 132:204–220.

Frankel, A. and Kartik, N. (2022). Improving Information from Manipulable Data. *Journal of the European Economic Association*, 20(1):79–115.

Hardt, M., Megiddo, N., Papadimitriou, C., and Wootters, M. (2016). Strategic Classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 111–122. Association for Computing Machinery.

Hu, L., Immorlica, N., and Vaughan, J. W. (2019). The Disparate Effects of Strategic Manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 259–268. Association for Computing Machinery.

John, L., Acquisti, A., and Loewenstein, G. (2011). Strangers on a Plane: Context-Dependent Willingness to Divulge Sensitive Information. *Journal of Consumer Research*, 37:858–873.

Kamel, H., Abdulah, D., and Al-Tuwaijari, J. M. (2019). Cancer Classification Using Gaussian Naive Bayes Algorithm. In *2019 International Engineering Conference (IEC)*, pages 165–170.

Kennedy, B., Funk, C., and 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. and Raghavan, M. (2020). How Do Classifiers Induce Agents to Invest Effort Strategically? *ACM Trans. Econ. Comput.*, 8(4):19:1–19:23.

Krishnaswamy, A., Li, H., Rein, D., Zhang, H., and Conitzer, V. (2021). Classification with Strategically Withheld Data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5514–5522.

Meir, R., Procaccia, A., and Rosenschein, J. (2012). Algorithms for Strategyproof Classification. *Journal of Artificial Intelligence*, 186:123–156.

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

Perez-Richet, E. and Skreta, V. (2022). Test Design Under Falsification. *Econometrica*, 90(3):1109–1142.

Pronk, V., Verhaegh, W., Proidl, A., and Tiemann, M. (2007). Incorporating user control into recommender systems based on naive bayesian classification. In *Proceedings of the 2007 ACM conference on Recommender systems*, pages 73–80. Association for Computing Machinery.

Sahu, S., Nautiyal, A., and Prasad, M. (2017). Machine Learning Algorithms for Recommender

System - a comparative analysis. *International Journal of Computer Applications Technology and Research*, 6:97–100.

Slokom, M., Hanjalic, A., and 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., and 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., and Lara-Cabrera, R. (2019). A Collaborative Filtering Approach Based on Naïve Bayes Classifier. *IEEE Access*, 7:108581–108592.

Wang, K. and Tan, Y. (2011). A new collaborative filtering recommendation approach based on naive Bayesian method. In *Proceedings of the Second international conference on Advances in swarm intelligence - Volume Part II*, pages 218–227. Springer-Verlag.

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., and Steinberg, D. (2007). Top 10 algorithms in data mining. *Knowl. Inf. Syst.*, 14(1):1–37.

A Appendix

A.1 Balance of characteristics in subject pools

Table 8 presents the characteristics of subject pools in *Info*, *Control* and overall. The first six lines give the frequency of answers to the binary questions of Part 1. The next lines give the proportion of subjects who had taken a course in statistics and average age. The last column presents p-values for t-tests of the differences between these characteristics in *Control* and *Info*.

Table 3: Characteristics of subjects in the *Control* and *Info* treatments

	Control	*Info*	Total	Diff. Control vs. Info (p-values)
ICE (% of Vanilla)	47.59	51.52	49.59	0.014
MUS (% of 3h+)	61.84	61.66	61.75	0.907
MAR (% of Yes)	54.30	60.24	57.32	< 0.001
NUC (% of Yes)	57.02	55.98	56.49	0.514
GEN (% of Male)	50.73	48.68	49.69	0.201
CHI (% of Yes)	53.04	54.56	53.81	0.341
Stats (% of Yes)	45.28	47.67	46.49	0.137
Age (mean)	41.92	41.55	41.73	0.376

A.2 Data of Part 3 of the experiment

Tables 4 (a), (b), (c) and (d) give the frequencies with which subjects gave each answer in Part 3. We remind the reader that, in Part 3, for each target question, we ask subjects to imagine they would have to guess a person's answer to that target question. We then ask them: *To make this guess, if you could see this person's answer to one other question, which one would be most useful?* They can choose between any of the five other questions and can also answer *None of the above questions would help me much to make that guess.* For example, Table 4 (a) reports what subjects answered when the target question was ICE. In *Control* and in *Info*, the most commonly-given answer (always in bold) is that no answer would help much to guess the answer to ICE (answer labelled 'NONE'). This answer is also the correct one (always in green).

The last columns of Tables 4 give p-values for t-tests of the differences between the frequencies of the various answers in *Control* and *Info* (significant differences appear in blue). In Tables 4 (a) and (b), we see that the *Info* treatment significantly decreases the share of subjects who answer that the target questions are correlated to no other questions. In Table 4 (b), we additionally see a significant increase in the share of subjects who answer that MUS is correlated to GEN, which correspond to the specific example we used when explaining subjects that the algorithm uses correlations. In Table 4 (d), we see that the *Info* treatment makes fewer subjects answer that NUC is correlated to no other question and

more subjects answer that NUC is correlated to GEN.

Table 4: Frequencies of answers in Part 3 for each target

	Control	*Info*	*Total*	*p-val*
NONE	**53.67**	**46.86**	**50.21**	0.034
CHI	10.06	9.74	9.90	0.865
GEN	29.77	35.50	32.68	0.057
MUS	3.14	3.25	3.20	0.929
MAR	2.10	3.45	2.78	0.201
NUC	1.26	1.22	1.24	0.954
Total	100	100	100	

(a) Freq. of answers in Part 3 for ICE(%)

	Control	*Info*	*Total*	*p-val*
NONE	**53.04**	**39.35**	**46.08**	<0.001
CHI	26.42	21.30	23.81	0.062
GEN	7.76	24.34	16.19	<0.001
ICE	2.10	2.23	2.16	0.886
MAR	9.01	11.56	10.31	0.193
NUC	1.68	1.22	1.44	0.549
Total	100	100	100	

(b) Freq. of answers in Part 3 for MUS(%)

	Control	*Info*	*Total*	*p-val*
NONE	10.69	9.13	9.90	0.415
CHI	**79.66**	**79.72**	**79.69**	0.984
GEN	5.24	5.07	5.15	0.905
ICE	0.84	0.81	0.82	0.963
MUS	2.94	3.25	3.09	0.780
NUC	0.63	2.03	1.34	0.058
Total	100	100	100	

(c) Freq. of answers in Part 3 for MAR (%)

	Control	*Info*	*Total*	*p-val*
NONE	**44.44**	**32.45**	**38.35**	< 0.001
CHI	27.46	31.03	29.28	0.222
GEN	20.96	31.64	26.39	< 0.001
ICE	1.05	1.22	1.13	0.804
MAR	4.19	2.03	3.09	0.052
MUS	1.89	1.62	1.75	0.754
Total	100	100	100	

(d) Freq. of answers in Part 3 for NUC (%)

A.3 Complementary regressions

Table 5: Uncorrelated targets (ICE and MUS)

	Optimal Strategy		
	(1)	(2)	(3)
Info	-0.123***	-0.122***	-0.124***
	(0.027)	(0.027)	(0.027)
Round		0.034***	0.034***
		(0.009)	(0.009)
Stats			0.054*
			(0.028)
Female			-0.024
			(0.027)
Age			-0.002
			(0.001)
Constant	0.463***	0.379***	0.432***
	(0.020)	(0.029)	(0.056)
Observations	1940	1940	1940

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

Table 6: Correlated targets (MAR and NUC)

	Optimal Strategy		
	(1)	(2)	(3)
Info	-0.007	-0.007	-0.008
	(0.022)	(0.022)	(0.022)
Round		0.004	0.004
		(0.009)	(0.009)
Stats			0.046**
			(0.022)
Female			0.038*
			(0.022)
Age			-0.002***
			(0.001)
Constant	0.282***	0.273***	0.329***
	(0.015)	(0.027)	(0.048)
Observations	1940	1940	1940

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

A.4 Selection of questions for the experiment and correlations

The selection of the six questions of Part 1 is tightly linked to the selection of the target questions. First, we wanted the experiment to be simple, so we wanted subjects to play the disclosing game with a small set of characteristics. Next, our goal was to find two target questions which answers were not correlated to any other answer given in Part 1, and two target questions which answers were correlated to only one (not more, for simplicity) of the answers given in Part 1. We also did not want the answers to the four target questions to be correlated to each other. Therefore, to pick the six questions, we computed the Pearson correlation coefficients for any pair of questions in the set of 30 questions that subjects completed in the pre-study. Table 7 gives these coefficients for the six questions we selected.[22] In the last column, the Table also presents the frequency of answers in the pre-study because we wanted answers that were relatively balanced a priori.

Table 7: Pearson correlation matrix for selected questions

	CHI yes	*MAR Yes*	*GEN Male*	*NUC Yes*	*MUS 3h+*	*ICE Vani.*	Freq.
CHI - Yes	1.00	0.47	-0.20	-0.10	-0.09	-0.09	0.47
MAR - Yes	0.47	1.00	-0.08	-0.05	0.02	-0.07	0.47
GEN - Male	-0.20	-0.08	1.00	0.29	0.10	0.09	0.51
NUC - Yes	-0.10	-0.05	0.29	1.00	0.05	0.08	0.51
MUS - 3h+	-0.09	0.02	0.10	0.05	1.00	0.02	0.64
ICE - Vani.	-0.09	-0.07	0.09	0.08	0.02	1.00	0.52

A.5 Proofs of Propositions 1 to 4

The proofs of Propositions 1, 2, 3 and 4 work by comparing the payoff obtained by the subject when disclosing a set D and the payoff obtained when disclosing a set D' which is strictly contained in D. It will be strictly beneficial for the subject to disclose D' instead of

[22] Section 1 in the Online Appendix gives the coefficients for the 30 questions of the pre-study.

D when

$$-2 * g_{D'} - 0.2 * |A \setminus D'| > -2 * g_D - 0.2 * |A \setminus D|$$

which can be rewriten as:

$$r(D, D') \equiv \frac{g_D - g_{D'}}{|A \setminus D'| - |A \setminus D|} > 0.1.$$

We have developed a procedure, named RATIO and detailed in section 4 of the Online Appendix, to compute the ratios $r(D, D')$ for various couples (D, D') and all possible sets A. In the proofs below, we detail how we use RATIO to identify subjects' optimal disclosure strategies.

Proof of Proposition 1. Consider a target question j and a subject characterized by A. The RATIO procedure first computes $r(D, D')$ for every $D \subseteq A$ such that $x_j \in D$ and every $D' = D \setminus \{x_j\}$. Next, RATIO repeats these computations for every possible A and for every possible target j. We find that the ratio $r(D, D')$ is always larger than 0.174. It means that, for every subject and target question, it is strictly beneficial to hide the answer to that target question. □

Proposition 1 implies that the optimal disclosure set never contains the answer to the target question j. In the three following proofs, we will compare subjects' payoffs for all disclosure sets that exclude x_j.

Proof of Proposition 2. Consider a target question $j \in \{ICE, MUS\}$ and a subject characterized by A. The RATIO procedure first computes $r(D, D')$ for $D = A \setminus \{x_j\}$ and $D' \subset D$. Next, RATIO repeats these computations for every possible A. We find that the ratio is always smaller than 0.050 for the target question ICE and always smaller than 0.051 for the target question MUS. It means that, for every subject and uncorrelated target

question, it is optimal to hide only the answer to that target question. □

Proof of Proposition 3. Consider a target question $j \in \{MAR, NUC\}$ and a common subject characterized by C^j. We call x_k the answer correlated to the answer x_j (the answer to CHI when j is MAR, and the answer to GEN when j is NUC). We will show that, for every common subject, it is optimal to disclose every answer except x_j and x_k. We call our candidate for the optimal disclosure set $D^* = C^j \setminus \{x_j, x_k\}$. We need to show that D^* dominates:

1. All disclosure sets that contain x_k (but still exclude x_j as prescribed by Proposition1). That is, the set $D = C^j \setminus \{x_j\}$ and all subsets $D' \subset D$ such that $x_k \in D'$.

2. All disclosure sets that exclude x_k, x_j and some other answers. That is, all subsets $D'' \subset D^*$.

While we can simply apply the RATIO procedure to prove 2, we have to proceed in two steps to prove 1. That is because our procedure is designed to compare a set with one (or more) of its subsets. However, all subsets $D' \subset D = C^j \setminus \{x_j\}$ such that $x_k \in D'$ are not subsets of D^* because they contain x_k. First, we show that the set $D = C^j \setminus \{x_j\}$ dominates all its subsets $D' \subset D$ such that $x_k \in D'$. That is, if x_k is disclosed, hiding only the answer to the target question dominates hiding the answer to the target question and any other set answers which does not contain x_k. Then, we show that our candidate $D^* = C^j \setminus \{x_j, x_k\}$ dominates the set $D = C^j \setminus \{x_j\}$. That is, hiding the answer to the target question and the correlated answer strictly dominates hiding only the answer to the target question. Then, by transitivity, we can conclude that $D^* = C^j \setminus \{x_j, x_k\}$ strictly dominates all disclosure sets which contain x_k.

The RATIO procedure first computes $r(D, D')$ for $D = C^j \setminus \{x_j\}$ and $D' \subset D$ with $x_k \in D'$. Next, RATIO repeats these computations for every possible C^j. We find that the ratio is always smaller than 0.035 for the target question MAR and always smaller than

0.051 for the target question NUC. It means that, for every common subject and correlated target j, if x_k is disclosed, hiding only the answer to the target question dominates hiding the answer to the target question and any other set answers. Second, the RATIO procedure computes $r(D', D^*)$ for $D' = C^j \setminus \{x_j\}$ and $D^* = D' \setminus \{x_j, x_k\}$. Next, RATIO repeats these computations for every possible C^j. We find that the ratio is always higher than 0.185 for the target question MAR and always higher than 0.120 for the target question NUC. It means that, for every common subject and correlated target, hiding the answer to the target question and the answer x_k strictly dominates hiding only the answer to the target question. As explained above, point 1 is finally proved by transitivity.

For point 2, the RATIO procedure computes $r(D^*, D'')$ for every $D^* = C^j \setminus \{x_j, x_k\}$ and $D'' \subset D^*$. Next, RATIO repeats these computations for every possible C^j. We find that the ratio is always smaller than 0.040 for the target question MAR and always smaller than 0.053 for the target question NUC, which proves point 2. We conclude that, for every common subject and correlated target question, it is optimal to hide both the answer to that target question and its correlated answer. □

Proof of Proposition 4. Consider a target question $j \in \{MAR, NUC\}$ and an uncommon subject characterized by U^j. The RATIO procedure first computes $r(D, D')$ for $D = U^j \setminus \{x_j\}$ and $D' \subset D$. Next, RATIO repeats these computations for every possible U^j. We find that the ratio is always smaller than 0.036 for the target question MAR and always smaller than 0.052 for the target question NUC. It means that, for every uncommon subject and correlated target question, it is optimal to hide only the answer to that target question. □

A.6 Discussion about the Naive Bayes Algorithm

In this section, we discuss the advantages of the Naive Bayes Algorithm (NBA). We first present its technical properties before explaining that it is widely used in practice.

As explained (see equation 1 in section 2.2), the NBA computes its guess $P(x_j|D)$ using

Bayes' rule and assuming that answers $\{\tilde{x_i}\}_{i \neq j}$ are mutually independent conditional on x_j. This has two important consequences. First, this makes every disclosed answer contributes independently to the guess of the algorithm. Indeed, disclosed answers contribute to the guess through the probability of the set D conditional on x_j, $P(D \mid x_j)$, which is simply the product $\prod_{x_i \in D} P(x_i \mid x_j)$. Without the independence assumption, $P(D \mid x_j)$ would correspond to the frequency (in the pre-study dataset) of the whole set D conditional on x_j; all answers in D would jointly determine the guess and it would be hard to assess the effect of a specific x_i. Second, the NBA can consistently handle missing data which is important as we allow subjects to hide their answers. Because each disclosed $x_i \neq x_j$ answer is treated independently, the NBA can simply exclude the hidden answers from the computation. This straightforward approach avoids the complexities introduced by other models, such as linear or logistic regressions, which require handling missing data through imputation (filling the missing answers) or recalculating the model with fewer variables. Overall, NBA creates a relatively intuitive setting for subjects to control their data, as noted by Pronk et al. (2007) and Valdiviezo-Diaz et al. (2019).

Naive Bayes is one of the most widely-used algorithm in machine learning. Companies use it widely used for classification tasks such as determining whether patients will develop cancers (Kamel et al. 2019) or filtering spams (Apache SpamAssassin or Mozilla Thunderbird). This algorithm is not the most commonly-used to perform recommendation tasks, which involve ranking rather than classifying, but it has been shown to perform well in some settings. For instance, Wang and Tan (2011) show that an improved version of the Naive Bayes algorithm performs better than the Amazon recommendation algorithm. Using the MovieLens dataset, a large set of individuals' movie ratings, Sahu et al. (2017) compare the ability of several algorithms to provide correct movie recommendations and conclude that the Naive Bayes approach is the most precise. According to the review by Wu et al. (2007), NBA ranks among the top 10 algorithms used in both the industry and the academic world for analyzing large datasets (data mining).

A.7 Additional pre-registered treatment - *Others*

We pre-registered three treatments, randomly assigned between subjects: *Control*, *Info*, *Others*. In the main text, we focus on the comparison between *Control* and *Info*. We now describe the *Others* treatment, and how it affects subjects' disclosure strategies compared to *Control*.

A.7.1 Description of the *Others* treatment

The objective of *Others* is to study whether subjects disclose information differently when they learn that disclosed information may be further used to train the algorithm. Precisely, in the *Others* treatment, subjects read the same sentences as in the *Control* treatment but we add the following text: *In subsequent experiments, we may use the answers you disclosed to further train our algorithm and make it better at guessing the answers of other individuals. This will be done in full anonymity as we record your answers anonymously.*

For this treatment, we had pre-registered the following hypothesis. It is based on the idea that subjects' behavior could be influenced by concerns about the future use of their disclosed information. Specifically, when subjects are explicitly told that their answers could be used to improve the algorithm's performance against others in future experiments, they may experience discomfort, guilt or hesitation about contributing to this process.

Hypothesis 3 *Pooling all target questions, subjects hide more answers in the Others treatment than in the Control treatment.*

A.7.2 Implementation

A total of 488 subjects were allocated to the *Others* treatment, each of them going through the four rounds of game against the algorithm. Considering the *Control* and *Others* treatments, we have 3860 observations in total.

The proportion of each answer given by subjects in Part 1 is not significantly different in *Control* and *Others*, except for slightly fewer subjects who do not have children and more subjects who prefer Vanilla ice cream in *Others*.

Table 8: Characteristics of subjects in the *Control* and *Others* treatments

	Control	*Others*	Total	Diff. Control vs. Others (p-values)
ICE (% of Vanilla)	47.59	51.02	49.32	0.033
MUS (% of 3h+)	61.84	63.11	62.48	0.415
MAR (% of Yes)	54.30	56.56	55.44	0.158
NUC (% of Yes)	57.02	55.33	56.17	0.289
GEN (% of Male)	50.73	50.20	50.46	0.743
CHI (% of Yes)	53.04	50.20	51.60	0.078
Stats (% of Yes)	45.28	51.02	48.18	< 0.001
Age (mean)	41.92	42.20	42.06	0.523

A.7.3 Analysis

We use the same experimental outcomes as in the main experiment, namely the frequency of optimal strategy and the number of hidden answers.

First, the frequencies with which subjects hide different number of answers is given in Figure 7. Over all observations, subjects hide on average 1.95 answers in *Others* and 1.88 in *Control* ($p = 0.142$). The distributions of these frequencies are not significantly different between *Control* and *Others* according to the Kolmogorov-Smirnov test ($p = 0.665$). This findings invalidate Hypothesis 3.

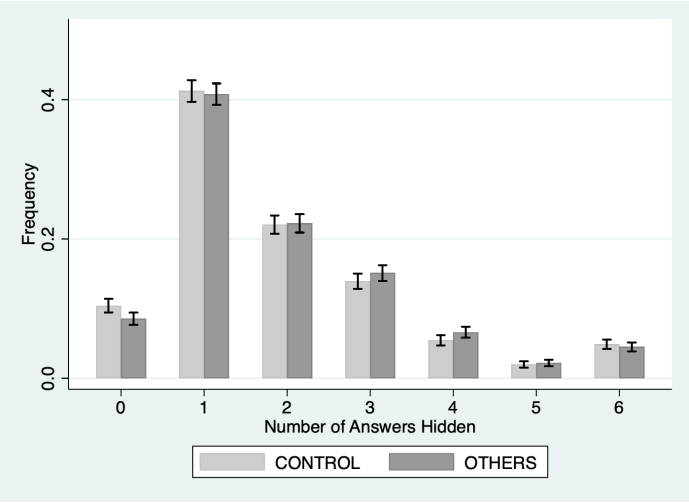


Figure 7: Hiding 0 to 6 answers, per treatment and pooling all targets

Second, over all observations, subjects play the optimal strategy 37.26% of the times

in *Control* against 36.48% in *Others*, which is not significantly different ($p = 0.612$). The absence of effect of the *Others* treatment compared to the *Control* is confirmed by the regressions presented in Table 9. When looking at each target question separately, we also observe no significant difference in the frequency of optimal strategy in *Control* and *Others* except for NUC (p-values for ICE, MUS, MAR and NUC are 0.924, 0.965, 0.631 and 0.024 respectively).

Result 6 *Pooling all target questions, there is no significant effect of the Others treatment compared to the Control treatment on the number of hidden answers or on the frequency of optimal strategy.*

Table 9: Optimal Strategies - All Targets

	Optimal Strategy		
	(1)	(2)	(3)
Others	-0.008	-0.008	-0.010
	(0.022)	(0.022)	(0.022)
Round		0.024***	0.024***
		(0.006)	(0.006)
Stats			0.043*
			(0.022)
Female			-0.002
			(0.022)
Age			-0.001
			(0.001)
Constant	0.373***	0.312***	0.331***
	(0.016)	(0.021)	(0.043)
Observations	3860	3860	3860

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