Strategic Information Disclosure to Recommendation Algorithms: An Experiment

Jeanne Hagenbach Aurelien Salas

June 27, 2024

Sciences Po Paris

Introduction

Recommendation algorithms

- Recommendation algorithms suggest or recommend products to individuals **based on their data** (past purchases, search history, demographics ...)
- These algorithms are trained on large datasets to identify **correlations between different pieces of information**
- From these correlations, algorithms **deduce non-disclosed information from disclosed information**
  - *Example:* A person who reveals having children has a good chance to be in a relationship
- By **strategically selecting what information to disclose or hide**, individuals can influence the algorithms' inferences and the recommendations they receive

Strategic Information Disclosure

Simple experiment

- Individuals' ability to **strategically disclose multi-dimensional information** to an algorithm when the objective is clear: **prevent the algorithm from guessing one particular trait**
- Investigate **what helps individuals achieve this goal**: information about the functioning of the algorithm? knowledge about correlations?

Contributions

- It is **complex** for individuals to manage their data, even in a very simple setting
- When subjects are informed of the algorithm's reliance on correlations, **they tend to hide more answers regardless of whether it is optimal to do so**
- To disclose information optimally, individuals must understand the functioning of the algorithm, the correlations known by the algorithm but **also the direction of these correlations**

Applied Relevance

User control and algorithm transparency are key regulatory issues

- **Transparency**
  - *GDPR Principle*: Individuals must be **clearly informed** about the collection and processing of their data (Articles 12-14)
  - *Assumption*: Individuals can **understand the information about the algorithms**
- **User Control**
  - *GDPR Principle*: Individuals have the **right to object to or restrict** data processing (Article 21)
  - *Assumption*: Individuals **can make disclosure decisions efficiently**

Our experiment highlights some of the challenges in designing transparent recommendation systems where users have control over their data

Literature (non-exhaustive)

- **Theoretical works** considering agents reporting private information to systems which decide allocations, attribute scores or classify users
  - *Economics:* design mechanisms/tests to reduce incentives to misreport - Eliaz and Spiegler (2019, 2022), Ball (2024), Frankel and Kartik (2022), Perez-Richet and Skreta (2022)
  - *Computer Sciences:* design algorithms robust to misreporting - Hardt et al. (2016), Krishanaswamy et al. (2021), Hu et al. (2019), Kleinberg and Raghavan (2019)
  - Miklós-Thal et al. (2023): disclosure of multi-dimensional data, hidden information can be deduced from disclosed information, study the dynamics of agents' disclosure

- **Experimental works** on how individuals disclose personal information online
  - *Psychology:* effect of various biasesSurvey in Acquisti et al. (2020); Instant gratification in Acquisti (2004); Herding in Acquisti et al. (2012); Status quo bias in John et al. (2011) ...
  - Bo, Chen and Hakimov (2023): experiment on manipulation of information transmitted by individuals to an algorithm doing personalized pricing

Experimental Design

Pre-Study

Before the main experiment, we collected answers from 505 *Prolific* participants to 30 binary questions (av. duration 2'51", fixed pay £0.6). Goal:

- Construct a short Study Questionnaire for the Main Experiment
- Train the algorithm

Experiment

The experiment has three parts. The instructions are given to the 970 *Prolific* subjects along the way (av. duration 8'43", av. pay £2.99)

- **Part 1:** Subjects complete a short Study Questionnaire
- **Part 2:** Subjects play the game against the algorithm
- **Part 3:** Post-game questionnaire

PART 1 : STUDY QUESTIONNAIRE

Subjects answer 6 questions, being told there is not right or wrong answer

- [MAR] Are you married or in a domestic partnership?
  Yes / No
- [CHI] Do you have children?
  Yes / No
- [NUC] Are you in favor of the use of nuclear energy?
  Yes / No
- [MUS] How much time do you spend listening to music per week?
  > 3 *hours* / ≤ 3 *hours*
- [GEN] What is your gender?
  Male / Female / Non-Binary
- [ICE] Which is your favorite ice cream flavor?
  Vanilla / Chocolate

- The questions are presented in one out of three random orders
- Subjects receive a fixed payment of $\pounds 1.2$ for Part 1

▸ Frequencies and Correlations

PART 2 : GAME AGAINST THE ALGORITHM

Subject plays **four rounds** of the game against the algorithm

In every round

- The algorithm **does not know the subject's answers to the Study questionnaire** (Part 1) but has been trained to guess them (using the data from the pre-study)
- The subject's objective is to **prevent the algorithm from guessing his/her answer to one *target question***
- The subject decides **which answers to hide from / to disclose to the algorithm** (no lies possible, hard info)

PART 2 : SCREENSHOT

The target question is : **Are you married or in a domestic partnership?**
Your task is to prevent the algorithm from guessing your answer was **Yes**.

Now you can decide which of your answers you want to disclose to the algorithm and which of your answers you want to hide.

Do you have children?

You answered Yes

Disclose this answer

Hide this answer

Are you in favor of the use of nuclear power?

You answered Yes

Disclose this answer

Hide this answer

How much time do you spend listening to music per week?

You answered 3 hours or less

Disclose this answer

Hide this answer

Which flavor of ice cream do you prefer?

You answered Chocolate

Disclose this answer

Hide this answer

What gender are you currently?

You answered Male

Disclose this answer

Hide this answer

Are you married or in a domestic partnership?

You answered Yes

Disclose this answer

Hide this answer

How does the algorithm work?

We use the Naive Bayes Algorithm, a commonly-used classification algorithm which

1. Learns from the pre-study dataset
   - **The prior probabilities** : frequency of each answer
   - For any two answers a and b to two different questions, the **probability of answer** a **conditional on** b : the frequency of answer a among subjects who answered b
2. Assumes answers are **independent conditional on the answer to the target**
3. Predicts the subject's answer to the target question using **Bayes' Rule**

How does the algorithm work?

- Our NBA algorithm generates **a probability distribution over the two possible answers to the target question**
- The **"guess of the algorithm"** is the probability estimated by the algorithm of the subject's **actual answer** given the information disclosed
    - In the example from the screenshot, the "guess" is the probability estimated by the algorithm that the subject answered YES to the question about marital status

▸ Formal definition

PART 2 : PAYOFFS

Trade off we focus on: hiding information is costly but may help not to be identified

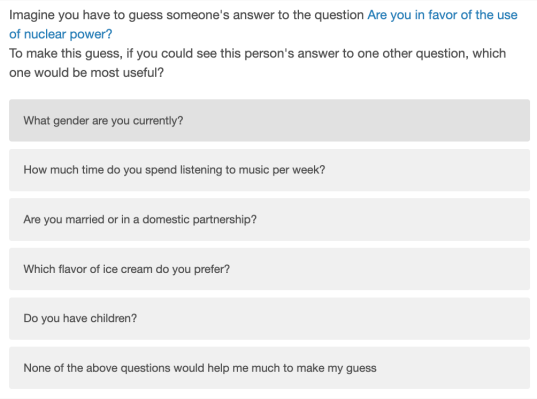
Subjects start Part 2 with an **endowment** of $£3.2$

- **Cost of hiding:** for each answer subjects decide to hide to the algorithm, we reduce the endowment by $£0.2$
- **Cost of "being identified":** Subjects' endowment is reduced more when the guess output by the algorithm is more accurate: it is reduced by $£2 \times$guess

With these payoffs and our NBA trained algorithm, **there always exists a unique, relatively simple disclosure strategy that maximises the subject's payoffs**

PART 3 : POST-GAME QUESTIONNAIRE

1. Incentivized questionnaire to elicit what subjects believe is the most correlated question to each target, or whether there is none (£0.10 per correct guess)



2. Socio-demographics questions: age, education level, whether they ever took a course in statistics

Experimental Treatments

What can make subjects play sub-optimally?

At least two possible reasons for subjects to play sub-optimally

- They **do not understand that the algorithm uses correlations**
- They understand that the algorithm uses correlations but wrongly estimate **which questions are most correlated with the target question**

We design two experimental variations to study these two possibilities

VARIATION 1 : INFORMATION ABOUT THE ALGORITHM

Text in the CONTROL treatment

What is the game about?

In every round of the game, you play against an algorithm. The algorithm does not know the answers you gave in Part 1 but it has been **programmed to guess these answers**.

In every round, your objective is to **prevent the algorithm from correctly guessing your answer to one specific question, the "target question".** Said differently, in every round, you must prevent the algorithm from learning one specific thing about you.

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.**

Text in the INFO Treatment: we add

To make this deduction, the algorithm has been trained on 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.

We implement this variation **between subjects**

VARIATION 2 : TARGET QUESTIONS

We selected four target questions to vary how obvious it is to identify correlations

MAR Are you married or in a domestic partnership?
Obvious correlation

NUC Are you in favor of the use of nuclear energy?
Non-obvious correlation

ICE Which flavor of ice cream do you prefer?
Uncorrelated to other questions

MUS How much time do you spend listening to music per week?
Uncorrelated to other questions

We implement this variation **within subjects**: every subject play with the four target questions in random order

Pre-Registered Hypotheses

Hypothesis 1 - Info Treatment

Given a target question, subjects reach the optimal strategy more often in the INFO treatment than in the CONTROL treatment

Hypothesis 2 - Correlations

Given a level of information about the functioning of the algorithm, subjects reach the optimal strategy more often when the correlation (or absence thereof) is easier to identify. Hence, subjects reach the optimal strategy less often when the target question is NUC than when it is MAR, ICE or MUS.

Optimal Strategies

Hiding the Answer to the Target

Hiding the Answer to the Target

It is always **strictly beneficial for the subject to hide the answer to the target question**

Intuition: Disclosing the target makes the algorithm guess the right answer for sure

Optimal Strategies - Uncorrelated Target Questions

Uncorrelated Target Questions

When **the target question is uncorrelated** (ICE or MUS), it is optimal for every subject to **hide only the answer to the target question**

Intuition : Hiding any other answer has a cost of $\pounds 0.20$ but (almost) doesn't reduce the guess of the algorithm

Optimal Strategies - Correlated Target Questions

For correlated target questions (MAR and NUC), the guess is largely determined by the answer to the question that is correlated to the target (CHI and GEN, respectively)

What did the algorithm learn from the Pre-Study?

Children $\Longrightarrow$ Married

Male $\Longrightarrow$ Pro Nuclear

Individually, subjects **may or may not** have answered like the majority of the subjects of the Pre-Study, and their optimal strategy will depend on that

- **Common subjects:** subjects who answered like the majority of subjects
- **Uncommon subjects:** subjects who have not answered like the majority of subjects
- These categories are defined for a given target question

Optimal Strategies - Correlated Targets

Correlated Target Questions - Common Subjects

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)

Intuition: Hiding the answer to the correlated question sufficiently decreases the algorithm guess relative to the cost of hiding

Correlated Target Questions - Uncommon Subjects

When the target question is correlated (MAR or NUC), it is **optimal for every uncommon subject to hide only the answer to the target question**

Intuition: The answers to the correlated questions ”mislead” the algorithm, disclosing them decreases the guess

RESULTS (IN PROGRESS)

DATA

- 970 subjects, each playing four rounds of game: **3880 rounds** / **disclosure choices**

	Uncorrelated (MUS & ICE)	Obvious Corr (MAR)	Non-Obvious Corr (NUC)	Total
Control	954	477	477	1908
Info	986	493	493	1972
Total	1940	970	970	3880

DATA

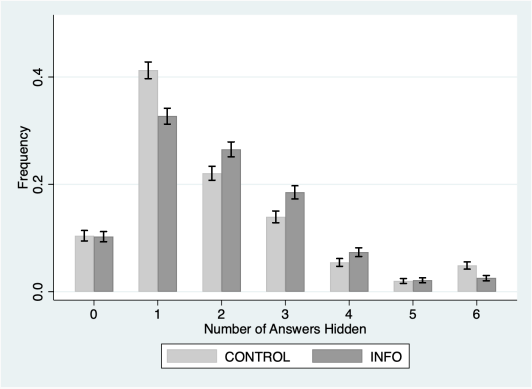
- Rationality check: subjects should never disclose the answer to the target question
  - In 79.95% of the rounds, subjects hide the answer to the target question
  - **63.30% of subjects hide the answer to the target question in all four rounds**
- 1.75% of subjects hide all answers in the four rounds
- 5.46% of subjects disclose all answers in the four rounds

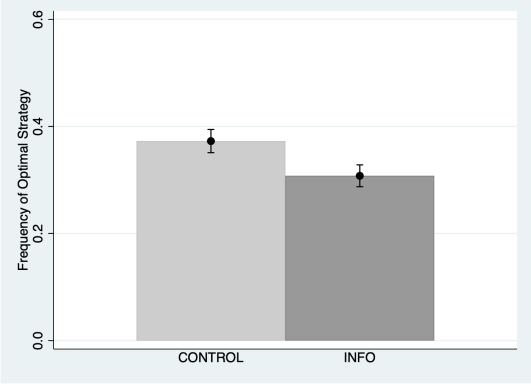
Most subjects seem to have understood the game and tried to play strategically

Main Outcomes of Interest

- Number of hidden answers, out of 6
- Whether or not the subject plays optimally
- Subject's beliefs about the most correlated answer to the target question

Overall Effect of Information





- Av number of answers hidden: 1.88 in CONTROL vs. 1.97 in INFO ($p = 0.064$)
- Distributions are significantly different (K-S : $p < 0.001$)

- Frequency of optimal strategy: 37.26% in CONTROL vs. 30.78% in INFO ($p < 0.001$)

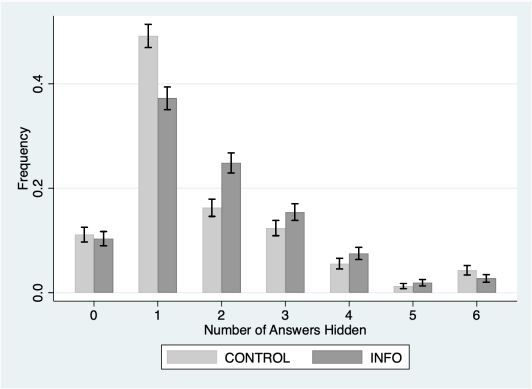
Overall Effect of Information

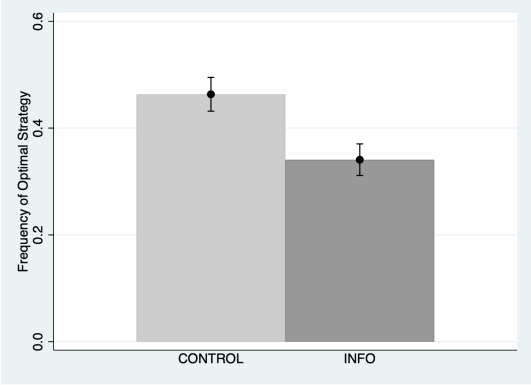
	Frequency of Optimal Strategy		
	(1)	(2)	(3)
Info	-0.065***	-0.065***	-0.096***
	(0.015)	(0.015)	(0.021)
StudiedStats			0.018
			(0.022)
Info * StudiedStats			0.064**
			(0.030)
Round			0.019***
			(0.007)
Constant	0.373***	0.459***	0.395***
	(0.011)	(0.027)	(0.034)
Demographics	No	Yes	Yes
Observations	3880	3880	3880

Note: The Table reports OLS coefficients (standard errors in parentheses). Demographics include age and gender.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

- The INFO treatment overall decreases the frequency with which players play optimal strategies
- INFO induces an overall increase in the number of hidden answers, which is sometimes sub-optimal
- Mainly driven by individuals who never studied Stats

Effect of Information - Uncorrelated





- Distributions are significantly different (K-S : $p < 0.001$)

- Frequency of optimal strategy: 46.33% in CONTROL vs. 34.07% in INFO ($p < 0.001$)

Effect of Information - Uncorrelated

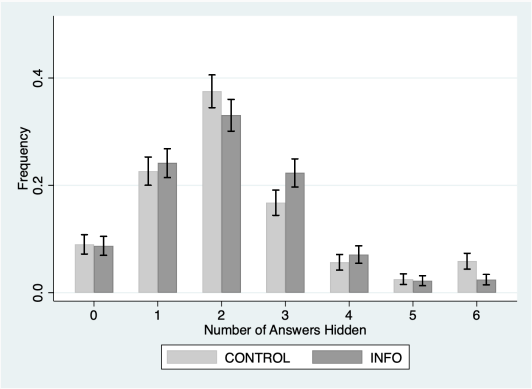
ICE&MUS	Frequency of Optimal Strategy (1)	(2)	(3)
Info	-0.123***	-0.124***	-0.168***
	(0.022)	(0.022)	(0.030)
StudiedStats			0.005
			(0.032)
Info * StudiedStats			0.095**
			(0.044)
Round			0.034***
			(0.010)
Constant	0.463***	0.550***	0.455***
	(0.016)	(0.039)	(0.049)
Demographics	No	Yes	Yes
Observations	1940	1940	1940

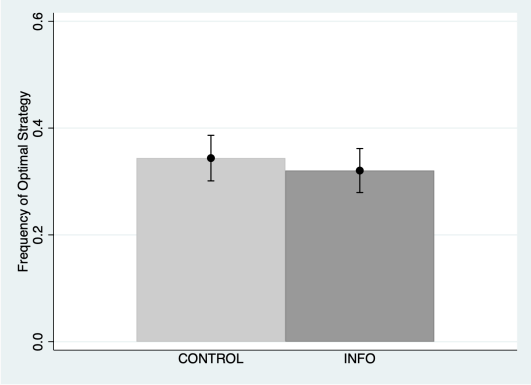
Note: The Table reports OLS coefficients (standard errors in parentheses). Demographics include age and gender.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

- For uncorrelated targets, the INFO treatment overall decreases the frequency with which subjects play the optimal strategy
- **The INFO treatment makes subjects *over-think***, look for correlations where there are none: detrimental when the optimal strategy is particularly simple
- Part 3: share of beliefs "none" decreases from 53% in CONTROL to 43% in INFO ($p < 0.001$)

▸ Beliefs

Effect of Information - Obvious Correlation





- Distributions are not significantly different (K-S : $p = 0.890$)

- Frequency of optimal strategy: 34.38% in CONTROL and 32.05% in INFO ($p = 0.441$)

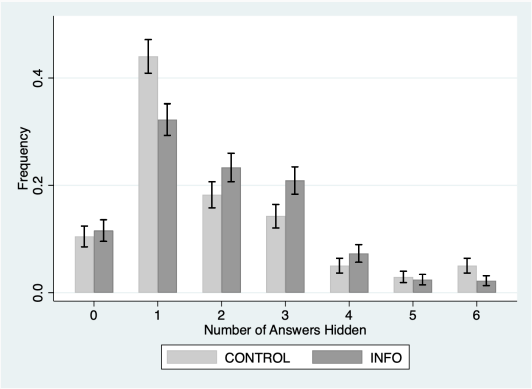
The Effect of Information - Obvious Correlation

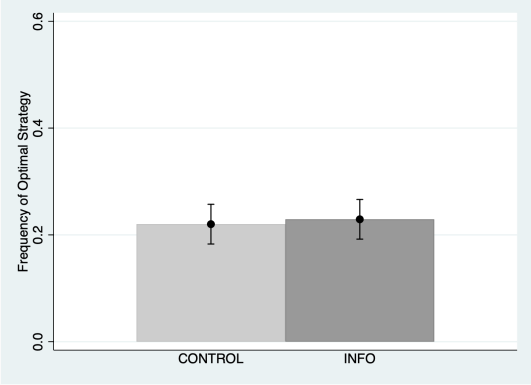
MAR	Frequency of Optimal Strategy		
	(1)	(2)	(3)
Info	-0.023	-0.025	-0.055
	(0.030)	(0.030)	(0.041)
StudiedStats			0.015
			(0.043)
Info * StudiedStats			0.064
			(0.061)
Round			-0.010
			(0.014)
Constant	0.344***	0.471***	0.484***
	(0.022)	(0.054)	(0.069)
Demographics	No	Yes	Yes
Observations	970	970	970

Note: The Table reports OLS coefficients (standard errors in parentheses). Demographics include age and gender.
$^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$.

- For the obviously correlated target, the INFO treatment has no significant effect on the frequency with which subjects play the optimal strategy
- **The INFO treatment does not help subjects to play better when correlations are obvious**
- Part 3: share of belief "CHI" is perfectly stable in CONTROl and INFO, 80% ($p = 0.98$) ▸ Beliefs

The Effect of Information - Non-Obvious Correlation

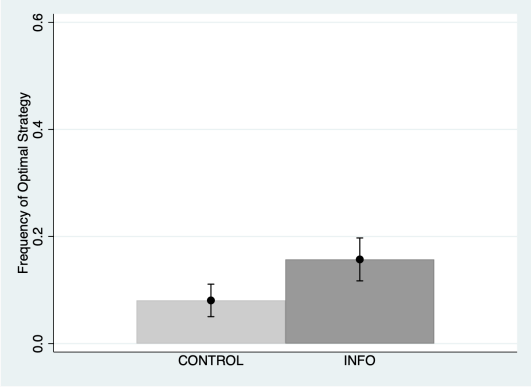




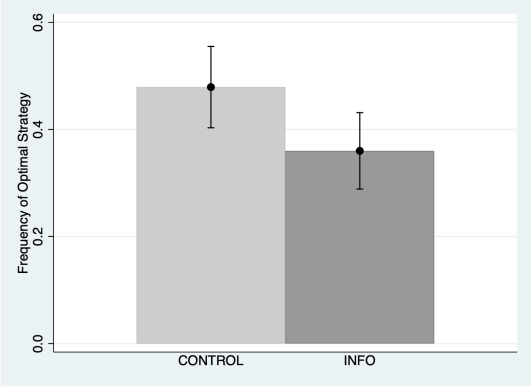
- Distributions are significantly different (K-S : $p = 0.008$)

- Frequency of optimal strategy: 22.01% in CONTROL vs 22.92% play it in INFO ($p = 0.735$)

The Effect of Information - Non-Obvious Correlation



- Frequency of optimal strategy for **common subjects**: 8.06% in CONTROL vs 15.72% in INFO ($p = 0.003$)



- Frequency of optimal strategy for **uncommon subjects**: 47.90% in CONTROL vs 36.00% in INFO ($p = 0.026$)

The Effect of Information - Non-Obvious Correlation

- For the non-obviously correlated target, the INFO treatment makes
  - Subjects identify significantly more often the correlation between NUC and GEN
  - Part 3: share of beliefs "GEN" increases from 21% in CONTROL to 32% in INFO ($p < 0.001$) Beliefs
  - 30.19% of subjects hide their answer to GEN in CONTROL vs 41.38% in INFO ($p < 0.001$)

The Effect of Information - Not Obvious Correlation - Common

NUC-Common	Frequency of Optimal Strategy		
	(1)	(2)	(3)
Info	0.077***	0.075***	0.056
	(0.026)	(0.026)	(0.035)
StudiedStats			0.001
			(0.037)
Info * StudiedStats			0.036
			(0.051)
Round			0.011
			(0.011)
Constant	0.081***	0.214***	0.184***
	(0.018)	(0.045)	(0.057)
Demographics	No	Yes	Yes
Observations	628	628	628

Note: The Table reports OLS coefficients (standard errors in parentheses). Demographics include age and gender.
$^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$.

- Hiding more in the INFO treatment significantly increases the frequency with which common subjects play the optimal strategy

The Effect of Information - Not Obvious Correlation - Uncommon

NUC-Uncommon	Frequency of Optimal Strategy		
	(1)	(2)	(3)
Info	-0.119**	-0.111**	-0.093
	(0.053)	(0.052)	(0.071)
StudiedStats			0.087
			(0.075)
Info * StudiedStats			-0.031
			(0.105)
Round			0.025
			(0.023)
Constant	0.479***	0.488***	0.374***
	(0.038)	(0.095)	(0.119)
Demographics	No	Yes	Yes
Observations	342	342	342

Note: The Table reports OLS coefficients (standard errors in parentheses). Demographics include age and gender.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

- Hiding more in the INFO treatment significantly decreases the frequency with which uncommon subjects play the optimal strategy
- **It is not enough to better identify the correlation, one needs to identify its direction**

Conclusion

Conclusion

A **simple disclosure experiment** to study whether individuals manage to strategically game a “recommendation algorithm” (the widely-used NBA)

- The easier to identify correlations are, the more frequently subjects play optimal strategies (Hypothesis 2 validated)
- The effect of information on subjects’ ability to play optimal strategies is ambiguous and heterogeneous (Hypothesis 1 not validated)
  - **Information make subjects overthink**, search for correlations even when there are none
  - Information may be enough to identify correlations but not necessarily to **identify the direction of these correlations**
  - **Information has heterogeneous effects** on subjects who studied stats and subjects who did not

Appendix

Frequencies of answers and Pearson Correlations in Pre-Study

	CHI Yes	*MAR Yes*	*GEN Male*	*NUC Yes*	*MUS 3h+*	*ICE Van.*	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 - Van.	-0.09	-0.07	0.09	0.08	0.02	1.00	0.52

Table 1: Pearson Correlation Matrix for Selected Variables

Back

FORMAL DEFINITION OF THE NBA

Environment:

- Six binary random variables ($\tilde{x}_1$ to $\tilde{x}_6$), each corresponding to a question.
- Each subject characterized by their six answers $A = \{x_1, x_2, x_3, x_4, x_5, x_6\}$.

Algorithm:

- Target question j; disclosed answers subset $D \subseteq A$.
- *Guess* $g_D \equiv P(x_j|D)$ computed using Bayes' Theorem, assuming independence of answers $\{x_i\}_{i \neq j}$ conditional on the answer to the target question x_j (Naive)

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

- With $P(D)$ given by:

$$P(D) = P(x_j) \prod_{x_i \in D} P(x_i|x_j) + P(\neg x_j) \prod_{x_i \in D} P(x_i|\neg x_j) \tag{2}$$

- Prior and conditional probabilities estimated from pre-study data via frequency analysis.

Special Cases:

- If the target question is disclosed, $g_D = 1$ (with Laplace smoothing, min 0.9829).
- If no answers are disclosed ($D = \emptyset$), $g_D = P(x_j)$.

Python package used: sklearn's BernoulliNB.

◂ Back

	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	0.9290
MAR	2.10	3.45	2.78	0.201
NUC	1.26	1.22	1.24	0.9542

Table 2: Distribution of beliefs in Part 3 for the ICE target (%)

	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

Table 3: Distribution of beliefs in Part 3 for the MUS target (%)

Back

PART 3 - BELIEFS FOR MAR TARGET QUESTION

	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

Table 4: Distribution of beliefs in Part 3 for the MAR target (%)

Back

PART 3 - BELIEFS FOR NUC TARGET QUESTION

	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

Table 5: Distribution of beliefs in Part 3 for the NUC target (%)

Back