

UNIVERSIT  D'ORL ANS
 COLE DOCTORALE SCIENCE DE LA SOCI T 
TERRITOIRE,  CONOMIE ET DROIT
LABORATOIRE D' CONOMIE D'ORL ANS

TH SE pr sent e par :

Mehdi LOUAFI

soutenue le : 9 mars 2026

pour obtenir le grade de : Docteur de l'Universit  d'Orl ans

Discipline/ Sp cialit  : Sciences  conomiques

FOUR ESSAYS ON HOMO NUMERICUS:
DECISION-MAKING IN THE AGE OF DIGITAL
CAPITALISM

TH SE dirig e par :

B atrice BOULU-RESHEF

Professeure des universit s, CY Cergy Universit  Paris,

THEMA

RAPPORTEURS :

Fabrice LE LEC

Professeur des universit s, Universit  de Lille, LEM

Marie Claire VILLEVAL

Directrice de recherche  m rite, CNRS, University of Lyon,
GATE

JURY :

B atrice BOULU-RESHEF

Professeure des universit s, CY Cergy Universit  Paris,
THEMA

Christophe HURLIN

Professeur des universit s, Universit  d'Orl ans, L O,
Pr sident du jury

Fabrice LE LEC

Professeur des universit s, Universit  de Lille, LEM

Matthias STEFAN

Associate Professor, University of Innsbruck, Depart-
ment of Banking and Finance

Jean-Christophe VERGNAUD

Directeur de recherche, CNRS, CES

Marie Claire VILLEVAL

Directrice de recherche  m rite, CNRS, University of
Lyon, GATE

L'université d'Orléans n'entend donner aucune approbation ni
improbation aux opinions émises dans les thèses ;
elles doivent être considérées comme propres
à leurs auteurs.

Remerciements

Au moment même où j'écris ces premières lignes, j'ai encore du mal à croire que ma thèse est sur le point de s'achever. Je dois, comme bon nombre de mes prédécesseurs, me confronter à cette fameuse épreuve des remerciements. Un sentiment de mélancolie m'envahit : malgré son lot d'inconvénients — solitude, doute, anxiété, fatigue et bien d'autres — cette thèse a aussi été, pour moi, un temps d'épanouissement. C'est donc avec une certaine nostalgie que j'en constate aujourd'hui la fin prochaine.

Mes premières pensées vont bien évidemment à ma très chère directrice de thèse, Madame Béatrice Boulu-Reshef, sans qui je n'aurais jamais pu débiter cette formidable aventure ni la mener à son terme. Béatrice, je tiens à vous remercier sincèrement pour la confiance que vous m'avez accordée. En arrivant au bout de cette thèse, je mesure pleinement la chance que j'ai eue de vous avoir à mes côtés, en tant que directrice, mentor, et même, au fil du temps, en tant qu'amie. Vous avez su m'inculquer les valeurs de la recherche, m'en faire comprendre les enjeux, mais surtout m'offrir l'opportunité d'entrevoir ce monde merveilleux et tout ce qui en fait une si belle vocation. Pour cela, je vous serai toujours profondément reconnaissant. Je garderai en moi de nombreux souvenirs de nos échanges : les digressions philosophiques, les discussions animées sur nos artistes et musiques préférés, les rires, les anecdotes... et surtout vos expressions. Des expressions que j'espère, un jour, avoir la chance de transmettre à mon tour à un-e jeune doctorant-e. Parmi tous les enseignements que vous m'avez transmis, le plus précieux est, à mes yeux, la place de l'humain dans la recherche. Vous avez toujours su le remettre au centre, y compris durant mes moments les plus difficiles. Je suis fier d'avoir mené cette thèse sous votre direction, et je vous remercie pour votre exigence, votre rigueur et votre bienveillance : cette thèse n'aurait pas été la même sans votre accompagnement.

Je voudrais également remercier Christophe Hurlin pour l'honneur qu'il me fait de présider le jury de cette thèse. Je remercie tout particulièrement Marie Claire Villeval et Fabrice Le Lec d'avoir accepté d'en être rapporteurs. Leur temps, leur expertise et la qualité de leurs remarques ont été précieux. Je suis également très reconnaissant à Jean-Christophe Vergnaud et Matthias Stefan d'avoir accepté de siéger au sein de mon jury. Leur participation à cette étape finale de mon parcours doctoral m'honore et m'est particulièrement appréciée.

Pour poursuivre ces remerciements adressés à des personnes exceptionnelles, je me dois de parler de celles et ceux qui composent le formidable laboratoire au sein duquel j'ai eu la chance de réaliser cette thèse et qui sont, au fil du temps, devenus une seconde famille. Les premières personnes qui me viennent à l'esprit sont bien sûr mes collègues et ami·e·s doctorant·e·s, avec qui j'ai partagé tant de choses. Marie, je te suis profondément reconnaissant pour notre si belle amitié : merci de m'avoir accompagné ces dernières années, d'avoir su m'écouter et me rassurer, d'avoir partagé avec moi ces fous rires, et, plus simplement, d'avoir été là. J'ai hâte de te voir grandir et de découvrir la chercheuse brillante et accomplie que tu deviendras. Flavien, merci pour ton enthousiasme, pour ton amour des quêtes secondaires et pour ces repas animés en salle de repos. Matilda, Sarah et Cristina, mes binômes de thèse : nous avons traversé cette épreuve ensemble, et je vous remercie sincèrement pour votre soutien. J'espère être à vos côtés lorsque nous serons tous docteurs, afin de pouvoir, ensemble, trinquer à ce succès. J'ai aussi une pensée pour tous les ancien·ne·s qui ont été, eux aussi, des repères précieux au début de mon doctorat. Wassim, Pauline, Inessa, Hugo, José, Indigo et Sébastien : merci de m'avoir transmis les ficelles du métier, pour nos discussions sur ce qu'est la thèse, ce qu'est la recherche, et, plus largement, pour avoir montré la voie. J'espère que ce travail portera, à sa manière, un peu de ce que vous m'avez appris. Vous me manquez, et je vous souhaite à toutes et tous de vous épanouir pleinement dans vos parcours respectifs. Enfin, je tiens à remercier West, Yannick, Nico, Aïssata, Ludivine, Martin, Aynur, Alexandra, Anne-Cécile, Julie et Linh, pour les moments partagés, la bonne humeur et la présence au quotidien.

Je ne peux pas continuer ces remerciements sans évoquer les maîtres de conférences et professeurs des universités qui m'ont guidé, parfois en quelques mots seulement, mais toujours avec justesse. Christophe, je vous remercie pour votre bienveillance et pour le rôle précieux que vous avez joué dans la dernière ligne droite de ma thèse. Alexis, travailler avec vous m'a énormément apporté, et j'espère que nos collaborations donneront lieu à de très belles publications. Anne et Raphaëlle, merci pour votre confiance : j'ai beaucoup appris à vos côtés lors de mes premières années en tant que chargé de TD, et je vous suis reconnaissant pour la bienveillance dont vous avez fait preuve à mes débuts. Et sur un registre plus personnel... Anthony... que dire ? Merci d'être le meilleur maître de conférences-doctorant que la Terre ait porté. Tu as toujours su être là pour nous écouter, nous conseiller et nous faire rire, dans les meilleurs comme dans les pires moments ; ma thèse n'aurait pas été la même sans toi. Yohan, ton arrivée a probablement été l'une des plus belles choses qui soient arrivées à ce labo et à moi aussi. Merci pour ta disponibilité, ton écoute, ta bienveillance et ton implication auprès des doctorants : que ce soit dans ton bureau, dans le mien, ou au détour d'une cigarette, nos discussions ont toujours été marquées par ta bonne humeur. Daria, il coccodrillo come fa ? Tu as été pour moi un éternel rayon de soleil : merci d'avoir su me faire rire à n'importe quel moment, et de m'avoir suivi dans mes délires les plus farfelus. Karine, je dois admettre que j'ai hésité à t'inclure dans cette partie, puisque tu es avant tout une amie à mes yeux (même si,

parfois, j'ai eu l'impression que c'était à sens unique...). Je tenais surtout à te remercier pour toutes les fois où la porte de ton bureau est restée ouverte, et où tu m'as écouté et rassuré. Merci aussi d'avoir supporté ce running gag qui dure maintenant depuis plus de trois ans. Gilles, merci pour tes blagues et merci d'avoir passé quatre ans à essayer de me faire comprendre ce qu'était l'économétrie non causale. Sache qu'à ce jour, je n'en ai toujours pas la moindre idée. Thaïs, merci de t'être toujours soucieuse de moi et de tes innombrables conseils glissés au détour d'un café. Théo et Adrien, même si vous êtes arrivés plus récemment au laboratoire, j'ai tout de suite apprécié pouvoir échanger avec vous : merci pour votre écoute, votre disponibilité et vos conseils toujours justes. Enfin, Françoise et Isabelle, merci pour votre aide précieuse, votre gentillesse et votre présence au quotidien.

Je suis également très reconnaissant envers le personnel administratif du Laboratoire d'Économie d'Orléans, dont l'aide et la disponibilité ont largement contribué au bon déroulement de cette thèse. Mélie, au-delà du fait que nous partageons deux dates d'anniversaire (le nôtre et bientôt celui de Clothilde), il y a bien d'autres choses qui nous lient. Je te remercie sincèrement pour ta bienveillance et ta gentillesse au quotidien, ainsi que pour ta présence et ton écoute, qui ont été précieuses pour moi. Merci aussi pour toutes ces fois où je t'ai appelé en panique à cause d'un problème administratif et où tu as su trouver une solution, toujours avec calme. Au-delà d'une collègue, tu es une véritable amie, et je te souhaite le meilleur pour la suite. Pierre-Yves, merci pour ta bonne humeur et pour avoir fait de moi ton habibi. Enfin, j'adresse une mention spéciale à Isabelle, Samuel, Émilie et Sandrine pour leur soutien, leur réactivité et leur aide tout au long de ces années.

Mon parcours doctoral m'a apporté non seulement un enrichissement académique, mais aussi des amitiés pour lesquelles je serai toujours profondément reconnaissant. Au fil de nombreuses conférences, j'ai eu la chance de rencontrer des personnes formidables et certaines ont su rester, année après année. Claire Mollier, merci d'être une personne exceptionnelle et d'avoir été, au fil du temps, ma "meilleure amie de la recherche". Tes innombrables retours ont profondément nourri mon travail, et je n'oublie pas notre objectif à peine ambitieux. Julien M'Barki, tu as toi aussi marqué cette thèse : je repense souvent au moment où nous avons eu la riche idée de commencer à coécrire ensemble... Merci pour cette aventure ; j'espère de tout cœur qu'elle se conclura par une belle publication.

Enfin, je voudrais adresser quelques mots à mes ami·e·s en dehors du monde académique, qui ont partagé — et parfois subi — ces années de thèse à mes côtés. Imane et Beya, je ne sais même pas par où commencer : merci d'avoir toujours été là, depuis plus de vingt ans. Merci d'avoir supporté l'éloignement que la thèse a créé, et de m'avoir soutenu sans relâche ; je vous aime fort. Romain, Léopold et Elsa, merci d'avoir été présents pour moi et d'avoir été, ici à Orléans, une véritable seconde famille. Alice, notre amitié est plus récente, mais tu as toujours su trouver la bonne manière de me changer les idées, de me faire prendre du recul et de remettre les choses à leur place. Skander et Yannis, vous êtes comme deux

frères : malgré la distance, vous avez toujours été là, et vous avez même pris le temps de vous intéresser à mes recherches et de m'en parler avec une curiosité qui m'a beaucoup touché. Juliette et Charline, merci pour votre présence et votre énergie, et pour tous ces moments qui m'ont permis de sortir un peu de la recherche et de respirer. Olessia, merci d'avoir été là dès le début de ma thèse à Orléans. Notre amitié a été une vraie force pour moi, et je suis très heureux qu'elle ait pris cette place dans ma vie.

Je suis éternellement reconnaissant envers ma famille. Même si des milliers de kilomètres nous ont séparés, votre amour et votre soutien ont toujours été présents et profondément ressentis. Je suis celui que je suis grâce à vous. Votre soutien indéfectible m'a permis d'avancer. À vous tous, je vous aime.

Enfin, je tiens à exprimer mes plus sincères remerciements à toutes celles et ceux qui, directement ou indirectement, ont contribué à mon parcours et m'ont aidé à atteindre cette étape importante. Votre soutien a été précieux, et je vous en suis profondément reconnaissant. J'ai également une pensée particulière pour celles et ceux qui se sont éloignés, brutalement ou en silence, comme cela arrive parfois. Ces quelques mots sont pour vous : vous demeurez présents dans ma mémoire.

Je voudrais conclure par deux dédicaces qui me tiennent particulièrement à cœur. À Daniel Cohen, d'abord, qui nous a quittés pendant ma thèse : son dernier livre a été un choc doux, une boussole et une source d'interrogations dont beaucoup se sont retrouvées, directement ou indirectement, dans ces pages. Et à Mark Knopfler, ensuite, dont la musique a été la bande-son de ce doctorat, des débuts hésitants jusqu'aux dernières relectures. À sa façon, elle a soutenu mon travail autant qu'elle a soutenu mes journées.

Merci à tous,
Mehdi

*"There's so many different worlds
So many different suns
And we have just one world
But we live in different ones"*
– Mark Knopfler

In memoriam Daniel Cohen (1953–2023).

Contents

List of Figures	vii
List of Tables	x
Introduction	1
1 Trading Returns for Privacy: Experimental Evidence from Financial Data Leaks	18
1.1 Introduction	18
1.2 Related literature	20
1.2.1 Definitions, subjectivity, and the economics of privacy	20
1.2.2 The privacy paradox and privacy calculus	21
1.2.3 Uncertainty, bounded rationality, and limits of the privacy calculus	22
1.2.4 Data leaks, financial information, and economic trade-offs	23
1.2.5 Social influences and individual heterogeneity in privacy decisions	24
1.2.6 Research questions and hypotheses	24
1.3 Experimental design and procedures	26
1.3.1 Allocation decision task	26
1.3.2 Experimental treatments	27
1.3.3 Implementation of data leaks	28
1.3.4 Questionnaires and additional tasks	29
1.3.5 Recruitment and data collection procedures	30
1.4 Results	30
1.4.1 Sample descriptives	30
1.4.2 Main trade-off between financial returns and privacy risk	31
1.4.3 Dynamic effects of data leaks	36
1.4.4 Individual heterogeneity	39

1.5	Conclusion	43
1.6	Appendix	46
1.6.1	Experimental Design	46
1.6.2	Additional Results	69
2	Algorithmic Transparency and Portfolio Choices: Field Evidence	73
2.1	Introduction	73
2.2	Related work	77
2.2.1	Explainable Artificial Intelligence	77
2.2.2	Explanations	78
2.2.3	XAI, explanations, and robo-advisors	80
2.2.4	Digital finance	81
2.3	Experimental design	82
2.3.1	Onboarding flow (“tunnel”)	82
2.3.2	Explanation Methodology	84
2.3.3	Outcome variables	86
2.4	Results	88
2.4.1	Descriptive statistics	88
2.4.2	Treatment Effect	90
2.5	Conclusion	99
2.6	Appendix	102
2.6.1	Experimental Design Appendices	102
2.6.2	Additional Results	103
3	Algo-Rhythm Unplugged: Effects of explaining algorithmic recommendations on music discovery	112
3.1	Introduction	112
3.2	Related work	115
3.2.1	Explainable Artificial Intelligence	115
3.2.2	Explanations	116
3.2.3	XAI, explanations, and music	117
3.3	Theoretical framework	118
3.3.1	Setup	118
3.3.2	Hypotheses	120
3.4	Experimental design	121

3.4.1	Behavioral Predictions	121
3.4.2	Experimental treatments	122
3.4.3	Questionnaire on music preferences and consumption habits	128
3.4.4	Tracks dataset	129
3.4.5	Laboratory Implementation	131
3.5	Results	132
3.5.1	Data and Empirical strategy	132
3.5.2	Treatment effect	134
3.5.3	Self-exposure to treatment	136
3.5.4	Model Predictions	139
3.6	Discussion	142
3.6.1	Engagement, interactions and music discovery behaviors	142
3.6.2	Treatment effect heterogeneity	144
3.6.3	External validity and Welfare perspectives	145
3.7	Conclusion	146
3.8	Appendix	150
3.8.1	Proofs	150
3.8.2	Experimental design appendices	155
3.8.3	Additional Results	168
4	Understanding, Trusting, Delegating: How Explanations Shape Public Support for	
	Algorithmic Decision-Making and AI Systems	181
4.1	Introduction	181
4.2	Related Literature	184
4.2.1	Explanations, transparency, and trust	184
4.2.2	Familiarity, experience, and belief formation	185
4.2.3	Delegation to algorithms and algorithm aversion in economic settings	186
4.3	Methodology	187
4.3.1	Survey Design and Items	187
4.3.2	ADM and AI systems	188
4.3.3	Sample and Data collection	188
4.4	Results	189
4.4.1	Descriptive Statistics	189
4.4.2	Empirical Strategy	192

4.4.3	Treatment Effect	195
4.4.4	Heterogeneity and additional analyses	200
4.4.5	Summary of results	208
4.5	Conclusion	209
4.6	Appendix	212
4.7	Survey design appendices	212
4.7.1	Items	212
4.7.2	Definitions and Explanations	214
4.8	Additional Results	219
4.8.1	Additional Figures	219
4.8.2	Descriptive Statistics	228
4.8.3	Additional Tables	239
General Conclusion		245
Bibliography		258
	References	259
Résumé in Extenso		275

List of Figures

1.1	Flowchart of the experiment	30
1.2	Average allocation choice in the risky option for personal data	32
1.3	Heatmap of average allocation to the risky option for personal data	33
1.4	Average allocation choice and allocation level	34
1.A.1	Decision Interface	46
1.A.2	Risk Aversion Elicitation Task à la Holt & Laury (2002)	47
1.A.3	Proportion of safe choices in each decision (Holt & Laury (2002))	48
1.A.4	Loss Aversion Elicitation Task à la Gächter et al. (2022)	48
2.1	Flowchart of the Experiment	83
2.3	Distribution of Recommended Risk Profile by Experimental Condition	88
2.4	Acceptance Rate per Recommended Risk Profile and Experimental Condition	89
2.A.1	Graphical explanation for the Risk Profile 10	102
2.A.2	Graphical explanation for the Risk Profile 6	102
2.A.3	Graphical explanation for the Risk Profile 3	102
2.A.4	Modification Page	103
2.A.5	Distribution of Devices Used by Experimental Condition	104
2.A.6	Distribution of Users' Type by Experimental Condition	104
3.1	Wording of the explanations presented to the treated group	125
3.2	Explanation diagram presented to the treated group	126
3.3	Number of tracks per music genre in the tracks dataset	130
3.4	Flowchart of the experiment	131
3.5	Relative listening time distribution and evolution across rounds	133
3.6	Average relative listening time (t_{ik}) by experimental condition	135
3.7	Self-exposure to treatment distribution and evolution across rounds by group	136

LIST OF FIGURES

3.8	Average end beliefs by experimental condition	140
3.9	Average end beliefs by self-exposure categories	141
3.A.1	Spotify’s “about recommendations”	155
3.A.2	Decision Interface	155
3.A.3	Distribution of Declared Genre Diversity by Declared Preferred Genre	156
3.A.4	Density Distribution of the Average Track Popularity per Playlist	159
3.A.5	Density Distribution of the Average Artist Popularity per Playlist	159
3.A.6	Genre proportions of generated Pop-oriented playlists	159
3.A.7	Average Relative Listening Time by group	168
3.A.8	Proportion of interactions by round and by self-exposure to treatment	169
3.A.9	End Beliefs: Post Attention Check Distribution	175
4.1	Decomposition per system, WTD1 (Weak)	198
4.2	Decomposition per system, WTD2 (Strong)	199
4.3	Item-level marginal gains of explanations	201
4.4	Conditional marginal gains of explanations by baseline system familiarity.	202
4.5	Treatment effects by political group	204
4.6	Attention to the explanation and willingness to delegate, pooled	205
4.7	Attention to the explanation and willingness to delegate, by system	207
4.A.1	Self-Reported Knowledge	219
4.A.2	Self-Reported Familiarity	220
4.A.3	Perceived Exposure	220
4.A.4	Self-Reported Use	221
4.A.5	Transparency Concerns 1	222
4.A.6	Transparency Concerns 2	222
4.A.7	Perceived Fairness 1	223
4.A.8	Perceived Fairness 2	223
4.A.9	Perceived Accountability 1	224
4.A.10	Perceived Accountability 2	224
4.A.11	Privacy Concern 1	225
4.A.12	Privacy Concern 2	225
4.A.13	Perceived Trust 1	226
4.A.14	Perceived Trust 2	226
4.A.15	Perceived Usefulness 1	227

4.A.16	Perceived Usefulness 2	227
4.A.17	Willingness to Delegate (Weak Delegation, WTD1)	228
4.A.18	Willingness to Delegate (Strong Delegation, WTD2)	228

List of Tables

1.1	Experimental design at a glance	26
1.2	Treatment overview	28
1.3	Determinants of Allocation Choices: Experimental Parameters	35
1.4	Determinants of Allocation Choices: Privacy Outcomes	37
1.5	Determinants of Allocation Choices: Privacy Outcomes (Cont.)	38
1.6	Determinants of Allocation Choices: Psychometric and self-reported measures	41
1.A.1	Acceptance Behavior by Lottery Choice Category	49
1.A.2	Preliminary questionnaire	49
1.A.3	Digital Knowledge and Privacy Measures	52
1.A.4	Instructions	68
1.A.5	Summary Statistics of Continuous Variables	69
1.A.6	Summary Statistics of Categorical Variables	70
1.A.7	Summary Statistics of Data included in Data Leaks	70
1.A.8	Consistency of responses between pre-registration and experimental questionnaire	71
1.A.9	Determinants of Allocation Choices: Economic and Behavioral Characteristics	71
2.1	Target asset allocation by risk profile (in %)	84
2.2	Explanations Effect on the Acceptance Rate	91
2.3	Explanations Effect on Deviations	92
2.4	Explanations Effect on Negative Deviations	93
2.5	CATE results for the Acceptance of the Recommendation	96
2.6	Characteristics of Users by CATE Group, Acceptance of the Recommendation	97
2.7	CATE results for the Deviation	97
2.8	User Characteristics by CATE Group and Method (Deviation)	98
2.A.1	Balance tests between treatment and control groups (non-parametric)	103
2.A.2	Summary Statistics for Behavioral and Engagement Outcomes	104

2.A.3	Number of observations per recommended risk level by experimental condition and in total.	105
2.A.4	Summary Statistics of Deviation by Recommended Risk Level and Experimental Condition	105
2.A.5	Explanations Effect on Positive Deviations	106
2.A.6	Explanations Effect on Negative Deviations (Cont.)	107
2.A.7	Explanations Effect on the Total Number of Attempts	108
2.A.8	Explanations Effect on the Total Number of Open Events	109
2.A.9	Explanations Effect on the Total Number of Modification Events	110
3.1	Behavioral Predictions Table	121
3.2	Conditions Overview	123
3.3	RSs Explanation Types	124
3.4	Summary statistics of the tracks dataset	131
3.5	Summary statistics of control variables and dependent variable	132
3.6	Explanations' effect on Relative Listening	136
3.7	Effect of self-exposure to treatment on Relative Listening	138
3.8	Effect of treatment and self-exposure on Model Predictions	143
3.9	Difference in mean and distribution of variables between self-exposure groups	145
3.A.1	Results of the Pearson's Chi-squared Test and Related Statistics	156
3.A.2	Preliminary questionnaire	160
3.A.3	Feedback questionnaire	165
3.A.4	Instructions	167
3.A.5	Summary statistics of numeric variables	170
3.A.6	Summary statistics of categorical variables	171
3.A.7	Balance test between experimental conditions	171
3.A.8	Effect of treatment and self-exposure on Likes and Dislikes	172
3.A.9	Effect of treatment and self-exposure on Jumps and Backs	173
3.A.10	Effect of treatment on end beliefs	174
3.A.11	Effect of self-exposure to treatment on end beliefs	174
3.A.12	Difference in mean of numeric variables between self-exposure to treatment groups	176
3.A.13	Distribution of categorical variables between self-exposure groups	177
3.A.14	Explanations' effect on Relative Listening (Poisson)	178
3.A.15	Effect of self-exposure to treatment on Relative Listening (Poisson)	179

4.1	ADM and AI systems studied in the survey	188
4.2	Treated vs Baseline: Welch two-sample t -tests and Wilcoxon rank-sum tests	189
4.3	Treated vs Baseline: Welch two-sample t -tests and Wilcoxon rank-sum tests	191
4.4	Treated vs Baseline: Welch two-sample t -tests and Wilcoxon rank-sum tests	191
4.5	Decomposition of Treatment Effects on Willingness to Delegate (Sum Indices)	196
4.6	Decomposition of Treatment Effects on Willingness to Delegate (PCA indices)	197
4.A.1	Initial Beliefs Items	212
4.A.2	Attitude and Willingness to Delegate Items	213
4.A.3	System types with definitions and explanations	214
4.A.4	Sample characteristics by country	229
4.A.5	Mean belief score per system and overall belief score, Baseline	231
4.A.6	Mean belief score per system and overall belief score, Treated	232
4.A.7	Mean attitude score per system and overall attitude score, Baseline (Panel A: Transparency, Fairness, Accountability)	233
4.A.8	Mean attitude score per system and overall attitude score, Baseline (Panel B: Privacy, Trust, Use)	234
4.A.9	Mean attitude score per system and overall attitude score, Treated (Panel A: Transparency, Fairness, Accountability)	235
4.A.10	Mean attitude score per system and overall attitude score, Treated (Panel B: Privacy, Trust, Use)	236
4.A.11	Mean WTD score per system and overall WTD score, Baseline	237
4.A.12	Mean WTD score per system and overall WTD score, Treated	238
4.A.13	WTD1 decomposition by system: Sum indices vs PCA indices	239
4.A.14	WTD2 decomposition by system: Sum indices vs PCA indices	241

Abstract

This dissertation studies *Homo Numericus*—individuals whose economic choices are increasingly mediated by digital platforms—and examines how algorithmic systems (recommendation, ranking, and personalization) shape decision-making and the collection of personal data. Across four experimental essays, I identify when and how privacy risks and transparency interventions through meaningful explanations affect behavior, beliefs, and acceptance of algorithms. Essay 1 quantifies the return–privacy trade-off in a laboratory investment task: allocations respond far more to return spreads than to leak probabilities; own leak experiences have small effects, while observing leaks among peers significantly reduces later privacy-risky choices. Essay 2 tests whether algorithmic transparency improves adherence in high-stakes financial decisions. The randomized controlled trial embedded in the onboarding interface of a robo-advisor (4,645 saving-plan contracts) shows that graphical explanations of the recommended risk score do not increase acceptance, reduce deviations, or raise engagement on average, but responses vary with users’ digital context and prior platform experience. Essay 3 shows that, in music recommendation, explanations improve beliefs about how the system works without a robust average change in listening behavior, with strong heterogeneity by attention to the explanation. Essay 4 finds that explanations generate small but detectable increases in willingness to delegate to algorithms. These effects arise primarily through improved attitudes toward the systems, while direct effects and belief shifts play a secondary role. Overall, transparency mainly shifts beliefs and attitudes: incentives, context of use, and social learning often dominate behavior. Implying that explainability and privacy policies should be tested empirically and tailored to domains and user contexts.

Résumé

Cette thèse étudie l'*Homo Numericus*, dont les décisions économiques sont de plus en plus médiées par des plateformes numériques, et analyse comment des systèmes algorithmiques (recommandation, classement, personnalisation) façonnent la prise de décision tout en collectant des données personnelles. À travers quatre essais expérimentaux, j'identifie quand et comment les risques de confidentialité et les interventions de transparence via des explications significatives influencent les comportements, croyances et l'acceptation des recommandations algorithmiques. Le premier essai mesure l'arbitrage rendement-fuite de données : dans une tâche d'investissement, les allocations réagissent davantage aux écarts de rendement qu'aux probabilités de fuite ; les fuites personnelles ont un effet faible, tandis que l'observation de fuites chez des pairs réduit durablement la prise de risque en matière de confidentialité. Le deuxième essai vise à déterminer si la transparence algorithmique améliore l'adhésion aux recommandations dans des décisions financières à forts enjeux. L'essai contrôlé randomisé intégré à l'interface d'un robot-conseiller (4 645 contrats), montre que des explications graphiques du score de risque n'augmentent pas l'adhésion ni l'engagement en moyenne, mais que les effets varient avec le contexte numérique et l'expérience des utilisateurs. Le troisième essai montre que, dans la recommandation musicale, les explications améliorent les croyances sur le fonctionnement du système sans effet moyen robuste sur l'écoute, avec une forte hétérogénéité selon l'attention portée aux explications. Le dernier essai montre que les explications génèrent une augmentation faible mais détectable de la volonté de déléguer à des algorithmes. Ces effets s'expliquent principalement par une amélioration des attitudes envers les systèmes, tandis que les effets directs et les changements de croyances jouent un rôle secondaire. Au total, la transparence et les explications modifient surtout les croyances et les attitudes : incitations, contexte d'usage et apprentissage social dominant souvent les comportements, ce qui plaide pour des politiques d'explicabilité et de confidentialité évaluées empiriquement et adaptées aux usages.

Introduction

Introduction

Decision-making in the age of digital capitalism

A growing share of economic life now takes place inside digital environments: we invest through robo-advisors, discover music through recommender systems, purchase through ranked lists, and interact with public and private services through interfaces that predict and pre-fill our preferences. These environments do not merely *display* information; they actively shape what is seen, in what order, and how it is justified. Ranking and personalization systems determine what enters the decision-maker's consideration set, while data collection turns each click, scroll, and choice into information that can be used to predict and sometimes influence future behavior. This dissertation studies decision-making in this context by focusing on *Homo Numericus*: individuals whose economic choices are increasingly mediated by digital environment and algorithmic systems.

The central premise is that once algorithms recommend, rank, and personalize, the relevant *information structure* faced by individuals changes. Classical information economics emphasizes that incentives and outcomes depend crucially on what agents know, what they believe, and how information is distributed and revealed (Stigler, 1961; Akerlof, 1970; Spence, 1973; Stiglitz, 1975; Grossman & Stiglitz, 1980). Digital intermediation changes that structure in two directions at once. First, the algorithm becomes an information producer for the user (scores, rankings, recommendations). Second, the platform becomes an information extractor from the user (data traces, behavioral signals), turning personal information into an input and an asset (Acquisti, Taylor, & Wagman, 2016; Srnicek, 2017; Zuboff, 2019). This dynamic is particularly visible in consumer finance, where machine-learning models increasingly exploit new digital data sources for risk assessment and credit scoring, raising issues of information use, governance, and potential bias (Hurlin & Pérignon, 2019). As a result, the economics of decision-making increasingly depends on how people understand, trust, and respond to systems that are simultaneously advising them and learning from them.

This transformation is particularly visible in *platform-mediated* markets. Platforms often match distinct user groups and monetize attention, data, or transactions. In the two-sided markets framework,

platforms are not passive venues; they design prices, rules, and interfaces that influence participation and behavior (Rochet & Tirole, 2003, 2006). Modern recommender systems can be seen as an additional layer of platform design: they implement a form of *algorithmic intermediation* that shapes attention and choice within the platform. Whether this intermediation improves welfare depends on the platform’s incentive objective function and constraints and on how users process the information they receive.

Two frictions lie at the heart of the dissertation. The first is privacy risk as an attribute of the choice set. The benefits of digital services (convenience, personalization, and lower search costs) are often bundled with data collection and exposure to misuse of personal information or leaks. The second friction is opacity and informational asymmetry in algorithmic systems. Users rarely observe how an algorithm operates, what objective it optimizes, and how their data enters the mapping from inputs to outputs (Eslami et al., 2015). They must decide whether to follow recommendations and whether to delegate decisions under imperfect knowledge.

In response to the second friction, regulators and firms increasingly promote transparency interventions: explanations, disclosures, and “meaningful” information about automated systems. In the European Union, this vocabulary is now explicit in the AI Act, which grants affected persons a right to obtain clear and meaningful explanations of the role of a high-risk AI system in an individual decision-making procedure (Article 86; Recital 171, European Parliament and Council of the European Union (2024)). Yet it is not obvious that explanations can reliably change behavior, trust, or acceptance. Explanations may update beliefs while leaving choices unaffected; they may work for some users and not others; and they may be dominated by incentives, default paths, and attention constraints. The dissertation, therefore, treats transparency and explainability not as normative endpoints, but as empirical interventions whose effectiveness can be measured.

Across four essays, I combine laboratory experiments, a randomized controlled trial, and large-scale online experiments to identify when and how privacy risks and transparency interventions through meaningful explanations shape behaviors, beliefs, and acceptance of algorithms. The thesis’s main finding is consistent across contexts: transparency and explanations can influence beliefs, but they are not a universal lever for changing behavior or acceptance. Incentives, context of use, and social learning often dominate.

From *Homo Economicus* to *Homo Numericus*. Economics traditionally models agents as decision-makers facing constraints and uncertainty. Two broad refinements have enriched this depiction. First, information economics formalized how asymmetric information and imperfect signals shape markets and contracts (Stigler, 1961; Akerlof, 1970; Spence, 1973; Stiglitz, 1975; Grossman & Stiglitz, 1980). Second, behavioral economics documented systematic departures from full rationality, such as bounded rationality, heuristics, and limited attention (Simon, 1955; Tversky & Kahneman, 1974; Kahneman & Tversky, 1979;

Gigerenzer & Goldstein, 1996; Gigerenzer & Gaissmaier, 2011). Digital capitalism combines and amplifies both themes.

Algorithmic systems intervene directly in the decision process. At the *menu* level, they select and order options, shaping consideration sets, perceived variety, and search costs. At the *belief* level, they generate signals and narratives about what is relevant, safe, popular, or optimal. A user choosing a savings plan, for example, does not face the full space of portfolios; she faces a recommendation, a restricted set of deviations, and an interface that frames risk and return. A listener exploring music does not face the universe of tracks; he faces a ranked list of suggestions and a “discover” flow that concentrates attention. In both cases, the algorithm is not merely a tool; it alters the choice environment by restricting and ordering the choice set and by shaping the signals on which individuals condition their decisions.

At the same time, platforms extract information from behavior: clicking, searching, and consuming become data. In many markets, this data improves matching and personalization; it can also enable persuasion, discriminatory pricing, or exposure to privacy harms (Acquisti et al., 2016). From the user’s perspective, the decision problem becomes layered. The user chooses content, products, or portfolios, but also implicitly chooses what information to reveal and what degree of delegation to accept. This is one reason why *Homo Numericus* is a useful label: it describes a decision-maker whose preferences are filtered through algorithmic mediation and whose actions are also informational signals.

The label captures several intertwined features. Choices are mediated by recommender systems, rankings, and predictive tools that filter the effective choice set and the information that becomes salient. At the same time, decisions are data-dependent: individual actions generate traces that feed back into the system and condition future recommendations. This interaction is characterized by opacity and informational asymmetry, as users rarely observe the system’s objectives, constraints, or internal logic. Finally, outcomes are context-sensitive because attention and deliberation depend on interface design and on the conditions under which information is encountered.

This dissertation does not claim that standard economic modeling becomes obsolete. Instead, it argues that algorithmic mediation is an information technology in the precise sense of information economics: it changes signals, beliefs, and the mapping from preferences to choices. Empirically, this implies that the welfare consequences of algorithmic systems cannot be inferred from technical performance alone. They depend on how users perceive risk, process explanations, update from experience, and decide whether to delegate.

Algorithms as information intermediaries: signals, choice architecture, and information design

A useful way to interpret algorithmic mediation is through information processing. In classical models, agents face uncertainty and choose actions based on signals. Once attention is limited, however, the *format*, *salience*, and *cognitive cost* of information become economic primitives (Sims, 2003; Chetty, Looney, & Kroft, 2009; Matějka & McKay, 2015; Gabaix, 2019). Digital interfaces directly shape these primitives: they determine what is salient, what is hidden behind clicks, and how costly it is to acquire additional information.

Algorithms as signals and as choice architecture. First, algorithmic outputs are signals: recommendations convey information (or at least *appear* to convey information) about match quality, risk, popularity, or social endorsement. Users may treat these outputs as expert advice, a summary statistic, or a persuasive cue. The literature on algorithm aversion and appreciation highlights that responses to algorithmic advice depend not only on accuracy but also on perceived competence, fairness, and the salience of errors (Dietvorst, Simmons, & Massey, 2015; Logg, Minson, & Moore, 2019). In other words, the same recommendation can be interpreted differently depending on beliefs about the system and prior experience.

Second, algorithmic interfaces are choice architecture: they curate menus and frame options, sometimes improving outcomes, sometimes biasing behavior (Thaler & Sunstein, 2008). In a robo-advisor, the recommendation is not merely information; it is the default path through onboarding. In a streaming platform, the recommendation list is not merely a forecast of taste; it is the main gateway to consumption. Thus, the economic role of algorithms is simultaneously informational and institutional: they are both signals and mechanisms that allocate attention.

This dual role highlights why it is useful to distinguish *understanding* from *behavior*. Users may come to understand what an algorithm “uses,” yet still follow it because the interface makes deviation costly; conversely, users may deviate even if they understand the system, if they perceive misalignment or if the stakes are high. The dissertation’s empirical designs are constructed to measure both.

Algorithms as persuasion and information design. A further layer is that algorithmic systems can be designed not only to predict, but also to influence. Platforms may have incentives to steer behavior toward higher engagement, higher margins, or lower risk exposure. From an economics perspective, this can be connected to information design and persuasion: the sender (platform/algorithm) chooses what signals to provide to the receiver (user) to shape actions (DellaVigna & Gentzkow, 2010; Kamenica & Gentzkow, 2011; Bergemann & Morris, 2019). Importantly, persuasion does not require deception; it can

arise through selective disclosure, framing, and ranking.

This perspective matters for transparency. If an explanation is itself part of the communication policy of the sender, then explanations can be strategic: they may reveal some aspects of the algorithm while obscuring others. Users may also anticipate such a strategy and discount explanations accordingly. This tension motivates the principal-agent framing developed below: explanations are attempts to reduce asymmetry, but they are delivered by an agent that may not be perfectly aligned.

Privacy as a risky attribute of options. Privacy introduces a distinct decision problem because it often involves probabilistic harms with uncertain magnitude and delayed consequences. Individuals trade immediate benefits (convenience, personalization) against uncertain future costs (leaks, misuse, discrimination) (Acquisti et al., 2016; Acquisti & Grossklags, 2005). In that sense, privacy can be modeled as a risky attribute bundled with the option itself.

A further complication is that privacy risk is inherently interpersonal in many digital settings. Consequences depend on who observes the information and how they react, and individuals may learn about risks not only through their own experience but by observing breaches affecting others. Such observational learning echoes mechanisms emphasized in information economics, including herding and informational cascades (Banerjee, 1992; Bikhchandani, Hirshleifer, & Welch, 1992). In digital contexts where severe events are infrequent for any single user but salient at the population level, vicarious learning can plausibly dominate direct learning. One contribution of the thesis is to quantify these channels in a controlled environment.

Transparency interventions: what explanations can (and cannot) do. In response to opacity and accountability concerns, regulators and firms increasingly promote transparency through explanations. The Explainable Artificial Intelligence (XAI) literature has developed diverse methods to interpret model predictions (Doshi-Velez & Kim, 2017; Adadi & Berrada, 2018; Arrieta et al., 2020), and widely used techniques such as LIME and Shapley-based decompositions have become reference points for producing local, prediction-level explanations (Ribeiro, Singh, & Guestrin, 2016b; Lundberg & Lee, 2017). Beyond explaining individual outputs, recent work proposes Shapley-based decompositions of a model’s overall predictive performance, which can help monitoring and accountability by identifying which features drive aggregate performance (Hu  , Hurlin, P  rignon, & Saurin, 2025). In high-stakes financial settings, transparency is often not merely an ethical aspiration but a practical and regulatory constraint, which has motivated “interpretable” machine-learning approaches that seek to retain predictive gains while preserving an intelligible decision logic (Dumitrescu, Hu  , Hurlin, & Tokpavi, 2022). Yet the economic impact of explanations depends on whether they meaningfully alter beliefs, behavior, and delegation.

Explanations may (i) increase perceived understanding, (ii) update beliefs about criteria and goals,

(iii) affect perceived fairness or accountability, and then (iv) change adherence or willingness to delegate. Alternatively, they may be ignored due to limited attention, misunderstood due to complexity, or even reduce trust if they highlight controversial trade-offs. Importantly, the effects of transparency can be heterogeneous because the cost of processing explanations differs across users and contexts. This is one reason why measuring engagement and not only providing the explanation is central in the dissertation.

A principal-agent view of algorithmic mediation

A central claim of this dissertation is that many interactions with algorithmic systems are naturally framed as a principal-agent problem. A principal delegates a task to an agent whose actions are imperfectly observable and whose objectives may not fully align with those of the principal (Ross, 1973; Jensen & Meckling, 1976; Holmström, 1979; Grossman & Hart, 1983). Delegation creates informational asymmetry and opens the door to moral hazard or strategic behavior when monitoring is limited.

Algorithmic systems, and the organizations deploying them, fit this structure in a direct way. Individuals delegate search, filtering, and evaluation to the system, and in some settings, high-stakes recommendations (e.g., portfolio risk profiles). This delegation occurs under opacity because users typically observe only outputs (i.e., scores or recommendations) and not the objective function, constraints, or training data that produce them. As a result, delegation is potentially accompanied by misalignment: the system may be optimized for engagement, monetization, retention, compliance, or risk management rather than the user’s welfare.

This lens becomes especially useful once we distinguish *ex ante* and *ex post* informational asymmetries. A distinction that structures the role of explanations throughout the thesis.

***Ex ante* asymmetry: deciding whether to delegate and disclose.** Before using an algorithmic system, individuals typically lack information about how the system uses their data, what it optimizes, and the distribution of outcomes they can expect. In principal-agent terms, the principal chooses whether to delegate without observing the agent’s type, incentives, or technology. This resembles adverse selection: users cannot fully distinguish systems that are competent and aligned from systems that are extractive or poorly governed.

In many markets, disclosure can partially unravel such asymmetries (Grossman, 1981; Milgrom, 1981). However, full unraveling is unlikely in algorithmic contexts: disclosures may be complex, attention may be limited, and firms may have incentives to provide minimal or strategic information rather than welfare-improving clarity (Gabaix & Laibson, 2006; McDonald & Cranor, 2008; Ben-Shahar & Schneider, 2014). As a result, *ex ante* delegation and disclosure (sharing data, accepting terms, onboarding) often occur under imperfect beliefs shaped by reputation, interface cues, and prior experience.

A related point is that *ex ante* opacity can interact with privacy. If data collection practices are difficult to observe, users cannot accurately price the privacy cost of delegation. This raises a classic information problem: when the quality dimension (privacy protection) is hard to observe, competition may not deliver the socially desired level of privacy, and disclosure requirements may be proposed as remedies. The dissertation provides evidence relevant to this logic by studying how individuals respond when privacy risk is made explicit and when algorithmic operation is explained.

***Ex post* asymmetry: deciding whether to follow recommendations.** Even after an individual adopts a system, a second informational problem arises: should the principal follow the agent’s recommendation? In finance, this becomes adherence to a recommended risk profile; in music, it becomes whether to accept recommendations or explore independently; across domains, it becomes willingness to rely on automated systems at all.

Ex post asymmetry matters because the principal cannot easily verify the basis of the recommendation or evaluate counterfactual alternatives. Communication between principal and agent is therefore central. In cheap-talk and strategic communication models, information transmission is constrained by incentives and credibility (Crawford & Sobel, 1982). If the sender’s objectives diverge from the receiver’s, messages may be partially informative at best. In algorithmic contexts, explanations can be interpreted as messages about the basis of recommendations. If users perceive misalignment, they may discount explanations; if they perceive alignment, explanations may increase reliance.

This is one way to understand why explanations might work in some settings and fail in others: their effect depends on baseline trust, perceived incentives, and the costs of verification. These conditions plausibly vary across domains (finance versus entertainment), across users (experienced versus novice), and across contexts of use (desktop versus mobile).

What does an explanation do in economic terms? This dissertation treats explanations as interventions that can affect outcomes through three channels. First, explanations can operate through an information channel by improving users’ understanding of what the system uses and why it produces a given output. Second, they can operate through an agency channel by affecting perceived alignment and trust: explanations may reassure users that the agent’s objective is compatible with their own, increasing delegation and adherence. Third, explanations interact with decision-making frictions through attention and cognitive costs. Because processing an explanation is costly, its effect depends on whether users allocate attention to it (Sims, 2003; Matějka & McKay, 2015).

A key empirical contribution of the thesis is to identify where this chain succeeds and where it breaks. Beliefs can move while behavior remains unchanged; behavioral effects can concentrate among users who actually engage with the explanation; and context (device, prior experience, stakes) can dominate. In

principal-agent terms, explanations may improve monitoring or perceived transparency, but this does not automatically change the effective contract if incentives, defaults, or switching costs keep behavior anchored.

Positioning in the literature: information, delegation, privacy, and transparency. The dissertation contributes to a fast-growing interdisciplinary literature, but its core questions are fundamentally economic: how do information structures created by platforms affect decision-making, and can transparency interventions mitigate informational asymmetry when individuals delegate to algorithmic systems in digital environments?

First, it builds on information economics, where search costs and information acquisition shape market outcomes (Stigler, 1961) and where the distribution and revelation of information are central to incentives and equilibrium (Akerlof, 1970; Spence, 1973; Stiglitz, 1975; Grossman & Stiglitz, 1980). Digital platforms reduce some search costs while introducing new informational frictions: users see curated signals but do not observe how they are produced.

Second, it connects to delegation and agency. Classic agency theory formalizes how hidden actions and hidden information shape contracts and organizational design (Ross, 1973; Holmström, 1979; Grossman & Hart, 1983). In algorithmic settings, the “agent” is a system embedded in an organization and a platform, and monitoring takes the form of transparency, interpretability, and governance. Strategic communication models emphasize that information transmission is limited by incentives and credibility (Crawford & Sobel, 1982), which provides a natural language for interpreting explanations as messages about algorithmic reasoning.

Third, it contributes to the economics of privacy, which emphasizes the trade-offs between benefits of data use and potential harms, and highlights why privacy choices may diverge from stated concerns (Acquisti & Grossklags, 2005; Acquisti et al., 2016). The dissertation’s first essay provides a benchmark environment where privacy risk is explicit and immediate, helping interpret behavior in real settings where probabilities are ambiguous and consequences are delayed.

Finally, it speaks to the emerging literature on explainability and acceptance of algorithmic systems, where explanations are expected to improve trust and legitimacy but evidence is mixed (Cramer et al., 2008; Doshi-Velez & Kim, 2017; Arrieta et al., 2020; Aysolmaz, Müller, & Meacham, 2023). By measuring belief changes, behavioral outcomes, engagement, and heterogeneity, the dissertation clarifies when explanations matter and when they do not.

Two core objects: privacy risk and meaningful explanations

The thesis focuses on two objects that lie at the center of digitally mediated choice and contemporary regulation: **privacy risk** and **meaningful explanations**. The expression “meaningful explanations” is used here in the spirit of recent EU transparency requirements for algorithmic systems, and the dissertation studies it empirically by asking whether explanations affect beliefs, behavior, and willingness to delegate.

Privacy risk: valuation, salience, and social learning. The privacy literature documents that individuals often express strong concern while disclosing substantial information in practice, a pattern sometimes described as a “privacy paradox” (Acquisti et al., 2016). One interpretation is that privacy preferences are unstable and context-dependent; another is that individuals face noisy beliefs and limited attention, while actual risks are ambiguous and consequences are delayed. Privacy can therefore be underweighted relative to immediate gains even when individuals care about it.

This dissertation contributes by studying privacy risk in a setting where probabilities are explicit, and outcomes are immediate: a dynamic investment task where one option offers a higher expected return but carries a known probability of a data leak in which selected financial information is revealed to peers. This design isolates a benchmark trade-off between money and privacy under transparent risk. A setting that helps interpret why privacy may be underweighted in the field, where probabilities are rarely known.

In addition, privacy beliefs may be shaped by vicarious experience. Users observe breaches affecting others, read about leaks, or infer risk from their peers’ outcomes. The thesis, therefore, pays particular attention to the role of social learning in privacy-related decision-making (Banerjee, 1992; Bikhchandani et al., 1992). This is not only descriptively relevant but also policy-relevant: if social learning is a major channel, then privacy interventions that rely on individual experience (or on quiet, individualized disclosures) may have weaker effects than interventions that make risks socially salient.

Meaningful explanations: a policy ideal with uncertain behavioral effects. Transparency requirements have proliferated across digital regulation and platform governance, often with the ambition of providing “meaningful” explanations: brief, accessible accounts of how systems work and what factors matter. From an economic perspective, explanations are intended to reduce informational asymmetry in principal-agent interactions, both *ex ante* (understanding the system’s logic and data use) and *ex post* (understanding the basis of a particular recommendation).

However, whether explanations translate into behavior change is not obvious. Explanations may improve perceived understanding and trust without altering choices, particularly when incentives, defaults, and time constraints dominate. Evidence on transparency and trust in recommender systems is mixed, and effects may depend on user goals and attention allocation (Cramer et al., 2008; Eslami et

al., 2015; Kizilcec, 2016; Miller, 2019; Poursabzi-Sangdeh, Goldstein, Hofman, Wortman Vaughan, & Wallach, 2021). More generally, disclosure can fail when information is complex or costly to process, even when it is technically “available” (Sims, 2003; McDonald & Cranor, 2008; Ben-Shahar & Schneider, 2014; Matějka & McKay, 2015; Gabaix, 2019). The thesis, therefore, evaluates explanations as interventions and measures their impacts on (i) beliefs, (ii) behavior, and (iii) acceptance and delegation.

Empirical strategy and roadmap of the dissertation

Why experiments? Experiments are well-suited for this agenda for three reasons. First, they allow direct manipulation of the information environment: leak probabilities, explanatory content, recommendation alignment, or disclosure format. Second, they help separate belief updating from incentive-driven behavior: one can observe whether explanations change understanding even when choices do not change. Third, experiments allow us to reveal mechanisms while taking into account heterogeneity, such as differential engagement with explanations or social learning from others’ outcomes.

The dissertation combines complementary designs:

- A laboratory setting with repeated privacy-risky choices and immediate social disclosure (Chapter 1),
- A field randomized controlled trial embedded in a real financial onboarding interface (Chapter 2),
- A controlled laboratory listening task studying discovery and engagement with explanations (Chapter 3),
- A large-scale representative online experiment measuring acceptability across algorithmic systems (Chapter 4).

A methodological theme is the joint measurement of **choices** and **beliefs**. Choices matter for welfare and policy, but beliefs are crucial for mechanisms. If explanations shift beliefs but not behavior, this suggests constraints or frictions (defaults, attention, incentives). If behavior changes without clear belief shifts, this suggests other channels (trust signals, perceived accountability, or simple salience). The dissertation’s designs are built to distinguish these cases.

A unifying conceptual thread. Although each essay stands alone, the dissertation is organized around a common logic: information about privacy risks or about algorithmic systems can update beliefs and attitudes, and these changes may translate into observed choices and willingness to delegate. This representation is deliberately stylized. Individuals start with prior beliefs, process information imperfectly under attention constraints, and their actions can generate experiences that feed back into later beliefs

and attitudes. The empirical objective is therefore to identify where the transmission from information to beliefs, and from beliefs to behavior, is strong or weak, and how it varies across contexts. In many digitally mediated settings, beliefs may shift while behavior remains anchored by incentives, defaults, or limited attention. The thesis, therefore, measures both belief outcomes and choice outcomes and documents heterogeneity across users and environments.

Roadmap of the dissertation: the four essays.

Chapter 1 - Trading Returns for Privacy: Experimental Evidence from Financial Data Leaks.

Question. How do individuals trade off expected financial returns against explicit probabilities of personal data leaks, and how do they update from their own versus others' leak experiences?

Design. In a dynamic laboratory investment task, one option offers a higher expected return but carries a known probability that selected financial information is revealed to peers; the alternative is privacy-safe. The task is repeated over many rounds, which permits the estimation of sensitivity to return spreads and leak probabilities while tracking dynamic learning from realized outcomes. The social nature of disclosure allows a clean distinction between (i) learning from one's own leak experience and (ii) learning from observing leaks experienced by others.

Main findings. Allocation choices respond much more to return spreads than to leak probabilities. Own leak experiences have modest and short-lived effects, whereas observing more leaks among peers significantly reduces subsequent choices of the privacy-risky option. In this environment, social learning about privacy outcomes is therefore a key driver of behavior.

Interpretation. Even under unusually transparent and immediate privacy risk, monetary incentives dominate. This benchmark helps explain why privacy-protective choices may be weak in natural markets where risks are opaque and delayed (Acquisti et al., 2016). The strong role of peer outcomes, moreover, suggests that privacy beliefs and behaviors may be shaped disproportionately by salient events and social information rather than by private experience.

Chapter 2 - Algorithmic Transparency and Portfolio Choices: Field Evidence.

Question. Does algorithmic transparency improve adherence in high-stakes financial decisions?

Principal-agent framing. Robo-advising is a canonical delegation problem: the user (principal) receives a risk profile recommendation from an algorithmic system (agent) whose mapping from inputs to outputs is opaque. Explanations aim to reduce informational asymmetry and improve perceived alignment, thereby increasing adherence and reducing costly deviations.

Design. In a randomized controlled trial embedded in the onboarding interface of a leading French robo-advisor, $N = 3,856$ users are randomly assigned to receive (or not receive) graphical, profile-based

explanations of the drivers of their recommended risk score. The trial covers 4,645 signed saving-plan contracts. The design focuses on outcomes that capture the strength of delegation: acceptance of the recommended profile, deviations from that recommendation, and engagement during onboarding.

Main findings. Providing explanations does not increase acceptance of the recommended profile, does not reduce deviations from the recommended risk profile, and does not raise engagement on average. Instead, behavior varies systematically with users’ digital context and prior experience (e.g., device and returning versus first-time status).

Interpretation. These results are consistent with a setting in which explanations can reduce opacity without altering the effective contract. If attention is scarce, the marginal value of processing an explanation may be low relative to simply proceeding. If users hold strong priors (about risk, about the platform, or about their own preferences), additional information may not shift behavior. Digital context matters because it shapes the cost of processing information and the friction of deliberation (Sims, 2003; Agnew & Szykman, 2005). From a broader perspective, the findings caution against assuming that transparency measures will automatically translate into better adherence or welfare-improving choices.

Chapter 3 - Algo-Rhythm Unplugged: Effects of explaining algorithmic recommendations on music discovery.

Question. Do explanations change users’ discovery behavior in recommender systems, and how does engagement with explanations mediate effects?

Design. In a 20-minute listening task, participants receive recommendations that are aligned with their stated preferences, misaligned with their stated preferences, or aligned with their stated preferences with an explanation of how the system works. The design measures both listening allocation and beliefs about the recommender’s criteria, and it records engagement with the explanation through time spent viewing it.

Main findings. Explanations improve participants’ beliefs about the recommender’s main criteria but do not produce a robust average change in relative listening time. Strong heterogeneity emerges: within the treated group, higher engagement with explanations (time spent on the explanation) is associated with more listening time and a shift in discovery behavior.

Interpretation. Explanations can work through belief updating, but behavioral effects depend on endogenous attention. This is consistent with limited attention and rational inattention: users benefit from explanations only if they choose to process them (Sims, 2003; Matějka & McKay, 2015). The results also highlight why “availability” of transparency is not equivalent to “use” of transparency. For policy and design, this suggests that explainability is partly an attention problem: how explanations are integrated into the user journey may matter as much as what they contain.

Chapter 4 - Understanding, Trusting, Delegating: How Explanations Shape Public Support for ADM and AI Systems in Europe.

Question. Do explanations causally increase the acceptability of algorithmic systems across domains?

Design. Using a representative sample from three large Western European countries (France, Germany, and Italy; $N = 3,007$), participants evaluate different algorithmic systems with or without an explanation of how each system works. Outcomes cover beliefs (knowledge, familiarity, exposure), attitudes (trust, fairness, privacy concerns, perceived usefulness), and willingness to delegate.

Main findings. Explanations yield context-dependent gains in baseline beliefs and in attitudes and willingness to delegate, with weaker or null effects in some domains. Overall, effects are modest and heterogeneous.

Interpretation. Explanations can shift acceptance, but they do not uniformly resolve concerns about privacy, fairness, and accountability. Concerns that are central to the governance of automated systems (Kroll et al., 2017; Aysolmaz et al., 2023). The context dependence suggests that users' priors and the perceived stakes of delegation shape how transparency is interpreted.

Synthesis: what the four essays jointly imply

Information about risk or algorithms $\rightarrow$ Beliefs and Attitudes $\rightarrow$ Choices and Delegation.

A central empirical conclusion is that the chain often breaks between beliefs and behavior. Explanations can be cognitively effective; they can improve stated understanding and sometimes attitudes without reliably changing what people do.

Beliefs may be changed more easily than choices. In the music experiment, explanations reliably improve understanding of the recommender's criteria, yet average listening allocation does not robustly change. In the cross-domain experiment, explanations modestly increase willingness to delegate, largely via attitudinal changes rather than large factual belief updates. In the robo-advisor field, explanations do not shift adherence on average. Together, these findings suggest that transparency is often more effective at shaping perceptions than at altering choices, especially when choices are anchored by defaults, constraints, or strong priors.

This pattern aligns with a simple attention-based interpretation. The cognitive cost of processing and integrating an explanation into a choice may exceed its marginal value, particularly when the recommendation is already plausible or when the decision must be made quickly. In rational inattention terms, explanations are information with a cost; whether they matter depends on whether the user optimally allocates attention to them (Sims, 2003; Chetty et al., 2009; Matějka & McKay, 2015; Gabaix, 2019).

The dissertation’s heterogeneity results (especially in the music task) reinforce this view by showing that effects concentrate among those who invest attention.

Incentives dominate privacy risk, even when risk is explicit. Chapter 1 provides a benchmark: even when leak risk is explicit and realized quickly, return spreads exert a stronger effect than leak probabilities. This does not mean privacy is irrelevant; rather, privacy risk is often second-order relative to immediate monetary incentives. This helps interpret why privacy protections that rely on individual choice (opt-outs, consent banners) may perform weakly when risks are opaque and delayed (Acquisti et al., 2016). It also suggests that making privacy salient may not be sufficient if the incentives embedded in the choice environment strongly favor privacy-risky options.

Social learning can outweigh personal experience. A striking result in Chapter 1 is that observing others’ leaks reduces subsequent privacy-risky choices more than experiencing one’s own leak. This resonates with information-economics models in which agents infer risk from others’ outcomes (Banerjee, 1992; Bikhchandani et al., 1992). In digital environments, where major breaches are rare for any single user but salient at the population level, social information may be a key driver of perceived risk and behavior. This has practical implications: privacy beliefs may react strongly to salient public events, and the behavioral consequences of privacy incidents may spill over to users who were not directly affected.

Digital context is an economic variable. Chapter 2 highlights that device and prior platform experience are not implementation details; they shape attention, deliberation, and the effective cost of processing explanations. Digital context, therefore, belongs in the information structure. Explanations encountered on a phone, during a rushed onboarding step, or by an experienced user may have very different effects than the same explanation encountered on a desktop by an inexperienced user. This suggests a broader lesson for digital policy: standardizing transparency *content* may not be sufficient if transparency is encountered under very different attention constraints.

Explanations as principal-agent tools: necessary but insufficient. Across essays, explanations function as partial remedies to informational asymmetry in principal-agent relationships. Yet principal-agent theory predicts that monitoring alone is rarely sufficient: incentives and constraints matter (Holmström & Milgrom, 1991). The dissertation’s results align with this view. Explanations can reduce opacity and move beliefs, but without complementary changes in incentives, governance, or accountability, they often have limited behavioral bite. In short, transparency can be helpful, but it is not a substitute for alignment.

Contributions to economics. The dissertation contributes to decision-making and information economics in four main ways.

A privacy benchmark under explicit, immediate risk. By providing a setting where privacy risk is explicit and realized immediately, Chapter 1 offers a clean benchmark for privacy valuation and learning. It identifies the dominance of monetary incentives and highlights social learning as a major channel shaping privacy-related choices. This complements broader work in privacy economics (Acquisti et al., 2016) by showing what happens when key field frictions: ambiguity, and delay are deliberately minimized.

Field-experimental evidence on explanations in high-stakes finance. Chapter 2 adds rare field evidence on explainability interventions in consequential household decisions. The null average effect is informative: it cautions against the assumption that explanations mechanically increase adherence and suggests that context and heterogeneity are central for evaluation. The findings also encourage a more agency-centered view of robo-advising: explanations can reduce informational asymmetry without changing behavior if other features of the delegated environment dominate.

Attention-mediated transparency effects in consumer platforms. Chapter 3 shows that explanations can shift beliefs without shifting behavior on average, while uncovering strong heterogeneity tied to engagement. This supports the view that explanation effects depend on endogenous information acquisition and attention constraints (Sims, 2003; Matějka & McKay, 2015). Methodologically, it motivates measuring engagement and interpreting transparency as a two-stage intervention: availability and consumption.

Cross-context causal evidence on acceptability and delegation. Chapter 4 provides cross-context causal evidence on how explanations shape attitudes and willingness to delegate. Domain dependence implies that one-size-fits-all transparency requirements may have uneven effects and should not be assumed to work uniformly (Aysolmaz et al., 2023). The results also offer a structured measurement approach to acceptability, separating beliefs, attitudes, and delegation.

Implications for policy and governance. Transparency requirements are often justified as consumer protection and accountability infrastructure. The dissertation supports their institutional relevance while cautioning against overpromising behavioral impacts.

First, transparency should be treated as an intervention to be evaluated empirically, not as a guarantee of better decisions. In many contexts, disclosure can fail due to complexity, limited attention, or strategic design (Grossman, 1981; Milgrom, 1981; Fung, Graham, & Weil, 2007; Gabaix & Laibson, 2006; McDonald & Cranor, 2008; Dranove & Jin, 2010; Ben-Shahar & Schneider, 2014). Second, tailoring matters: explanation design should reflect domain-specific concerns and the digital context in which information is encountered. Third, transparency is not a substitute for alignment. Even when explanations increase acceptance, they do not ensure that objectives are aligned or that harms are prevented; broader

accountability and governance remain necessary (Kroll et al., 2017).

A final implication concerns evaluation standards. Because explanation effects are heterogeneous, average effects can be misleading for policy. If only a subset of users engages with explanations (as in Chapter 3), then the relevant design problem becomes: how can transparency be integrated so that it is actually processed by those who need it? If effects depend on device and experience (as in Chapter 2), then explanation mandates should be assessed in the specific interface where they will live, not only on paper. In short, transparency should be assessed as part of an *information environment*.

Conclusion and dissertation structure. Overall, the dissertation offers an empirically grounded account of decision-making for *Homo Numericus* in the age of digital capitalism. It shows that explanations can reduce informational asymmetries in the principal-agent relationship between users and algorithmic systems, but that transparency is not a universal solution. Incentives, attention constraints, context of use, and social learning frequently shape behavior more strongly than explanations do. For economics, the implication is that algorithmic mediation should be treated as an information structure. Additionally, transparency and privacy policies should be evaluated empirically and tailored to domains and user contexts rather than assumed to work uniformly.

Chapter 1

Chapter 1

Trading Returns for Privacy: Experimental Evidence from Financial Data Leaks

1.1 Introduction

The digitalization of economic interactions has dramatically increased the amount of personal information that is collected, stored, and traded. Individuals routinely reveal detailed traces of their consumption, mobility, and social networks in exchange for access to online services or better prices. At the same time, data breaches and leaks have become a recurrent feature of the digital economy, with large-scale incidents regularly exposing the personal information of millions of individuals and imposing non-negligible costs on firms and consumers alike.¹ These developments raise the question of how individuals actually trade off economic benefits against the risk that their personal data may be disclosed.

Financial services provide a particularly salient context for studying this trade-off. Financial data – such as income, wealth, outstanding debt, and detailed payment histories – are among the most sensitive categories of personal information. They are central inputs into credit scoring, fraud detection, and risk management, and they increasingly feed into AI-driven decision systems (Boissay, Ehlers, Gambacorta, & Shin, 2021). Recent policy initiatives such as open banking and, more broadly, open finance seek to promote competition and innovation by facilitating consent-based data sharing between banks, fintechs, and other third parties (OECD, 2023). While these initiatives promise better financial products and improved

¹European Union Agency for Cybersecurity (ENISA) (2023), *ENISA Threat Landscape 2023*, ENISA, October 19, 2023. Available at <https://www.enisa.europa.eu/publications/enisa-threat-landscape-2023>.

inclusion, they also multiply the number of actors with access to granular financial data and, therefore, the potential attack surface for data breaches and misuse. Evidence from large credit-reporting and banking incidents suggests that financial data leaks affect individuals' behavior, for instance by inducing more frequent credit freezes and opt-outs from data sharing (Mikhed & Vogan, 2018). Understanding how individuals value the protection of their financial data, relative to financial returns, is thus crucial for assessing both the welfare effects of data-driven finance and the design of regulatory safeguards.

A large body of literature in economics and related fields has studied how individuals value privacy and how they behave when their data may be collected or shared. Theoretical and empirical work has emphasized the idea of a “privacy calculus,” in which individuals trade off expected benefits and perceived privacy costs, as well as the so-called “privacy paradox,” whereby stated concerns about privacy often coexist with seemingly lax disclosure behavior. This literature has documented the role of incomplete information, behavioral biases, and contextual framing in shaping privacy-related decisions. However, most experimental evidence focuses on relatively generic online data (such as browsing histories or social network profiles), often in static environments where privacy risks are only partially specified. Much less is known about how individuals behave when the data at stake are financial, when the probabilities of a data leak are transparent, and when decisions are repeated over time with feedback.

Most privacy calculus models implicitly assume transparent and well-defined probabilities over privacy outcomes, and decision makers who can weigh these probabilities against expected benefits. In practice, individuals often face severe uncertainty and ambiguity over whether and how their data may be leaked or abused, and rely on heuristics and boundedly rational strategies. Within such environments, observed choices may diverge sharply from stated concerns and intentions. In this paper, I study privacy decisions in a dynamic financial environment in which the probabilities of data leaks are made explicit to participants and leak outcomes are realized in real time. This choice is consistent with standard experimental protocols that avoid deception, and it also serves a methodological purpose: it allows us to isolate how individuals trade off financial returns and privacy risks when the relevant probabilities are clearly stated and stable. I view this setting as a benchmark, abstracting from some of the uncertainty and opacity that characterise real-world data environments.

In real-world data environments, individuals rarely know the likelihood, timing, or audience of a potential leak. To isolate the privacy calculus under well-defined risks, I study decisions in a benchmark laboratory setting in which leak probabilities are explicitly stated, and outcomes are realized immediately. Leaks are deliberately limited in scope and disclosed pseudonymously to other participants, so the paper speaks to behavior under transparent and immediate privacy risks rather than to a universal valuation of privacy across contexts.

I design a dynamic investment task in which participants repeatedly allocate an endowment between

a safe option and a risky option. The task is explicitly framed as an investment decision: in each round, subjects choose between two risky investment opportunities and decide how much of their endowment to invest in the selected option. Across rounds, the risky option varies along two dimensions: its expected financial return and the probability that choosing it will trigger a leak of the participant’s own financial data. Leaks consist in revealing to other participants a subset of the subject’s confidential financial information collected in a pre-experimental questionnaire (income, assets, debts, and other indicators). Leak probabilities are explicitly displayed, and outcomes (financial payoffs and data leaks) are realized immediately after each decision. This design allows us to separately identify how choices respond to financial incentives and to the risk of a data leak in a dynamic financial context with transparent and time-varying probabilities of real-time data leaks.

This paper documents three benchmark facts in this transparent-probability design. First, participants trade off returns and leak risk in the expected directions, but the response to return spreads is much larger than the response to leak probabilities, indicating an asymmetric sensitivity to financial versus privacy incentives in this setting. Second, the dynamic effects of leaks differ by source: own leaks have modest and short-lived impacts, whereas greater leakage among other participants is followed by a lower propensity to choose the privacy-risky option, consistent with a social-information (or “crowd”) channel. Third, standard economic preference measures (including risk and loss aversion) and most economic characteristics explain little of the heterogeneity in choices, while context-relevant privacy attitudes are more predictive. Taken together, the results clarify which behavioral patterns emerge when privacy risks are immediate, probabilities are explicit, and decisions are repeated with feedback.

The remainder of the paper is organized as follows. Section 2 discusses the related literature and derives testable hypotheses about the trade-off between financial returns and data leaks, the role of personal versus others’ leaks, and the impact of individual characteristics. Section 3 describes the experimental design and procedures. Section 4 presents the main empirical results on choices, dynamics, and heterogeneity. Section 5 discusses the implications of the findings for data-driven financial services and privacy regulation, and Section 6 concludes.

1.2 Related literature

1.2.1 Definitions, subjectivity, and the economics of privacy

A large interdisciplinary literature has emphasized that privacy is a multi-dimensional and fundamentally subjective concept. Philosophers, legal scholars and social scientists have alternately defined privacy as the protection of a personal sphere or “right to be let alone” (Warren & Brandeis, 1989), control over personal information and its disclosure (Westin, 1968), the regulation of boundaries between self and

others and between private, shared, and public domains (Altman, 1975), or as a precondition for dignity, autonomy, and individual freedom (Schoeman, 1992). Economic approaches have stressed the role of privacy in governing the flow of information in markets and organizations, with attention to strategic disclosure and concealment in personal and business contexts (Posner, 1978).

Building on these foundations, a specific field of privacy economics has emerged, distinct from classical information economics. While information economics typically examines the value of information per se, privacy economics focuses on information about identifiable individuals and on how such information is collected, processed, shared, and monetized (Acquisti et al., 2016). Because privacy concerns and attitudes are context-dependent and idiosyncratic, individuals exhibit a wide range of behaviors and strategies when deciding whether to share or protect data (Morando, Iemma, & Raiteri, 2014; Nissenbaum, 2004). This diversity reflects heterogeneous concerns about the accessibility of personal data, social and organizational threats, and the perceived sensitivity of different data types (Krasnova, Günther, Spiekermann, & Koroleva, 2009). This experiment contributes to this literature by focusing specifically on financial data, which previous work identifies as particularly sensitive and consequential for individuals’ economic opportunities (Mothersbaugh, Foxx, Beatty, & Wang, 2012; Carrascal, Riederer, Erramilli, Cherubini, & De Oliveira, 2013).

1.2.2 The privacy paradox and privacy calculus

Despite rising awareness of privacy risks, many empirical studies document a disconnect between stated privacy concerns and actual disclosure behavior, a phenomenon often referred to as the “privacy paradox” (Berendt, Günther, & Spiekermann, 2005; Norberg, Horne, & Horne, 2007). Survey and experimental evidence shows that individuals frequently express strong concerns about the misuse of their personal data while simultaneously engaging in behaviors that expose substantial amounts of information (Acquisti & Gross, 2006; Norberg et al., 2007; Turow, King, Hoofnagle, Bleakley, & Hennessy, 2009). Yet there is no consensus on whether this disconnect truly reflects inconsistent preferences or instead arises from measurement and contextual issues.

One influential line of work argues that the paradox may partly be an artifact of comparing broad attitudinal measures with highly specific behaviors. Attitudes are often elicited in general terms (e.g., “I am concerned about my online privacy”), whereas behaviors are observed in concrete, context-rich situations. As a result, weak or noisy correlations between concerns and behavior may not imply irrationality, but rather the limitations of global measures as predictors of situational choices (Acquisti et al., 2016). Other contributions emphasize that individuals may hold stable preferences but face complex trade-offs between privacy and other goals, leading to context-dependent decisions.

These trade-offs are captured by the notion of a *privacy calculus*, according to which individuals

mentally weigh anticipated losses of privacy against potential benefits of data disclosure (Laufer & Wolfe, 1977; Culnan & Armstrong, 1999; Dinev & Hart, 2006; H. J. Smith, Milberg, & Burke, 1996). Empirical work has documented such trade-offs in a variety of settings, showing that individuals are willing to disclose more data in exchange for price reductions, personalization, convenience, or social benefits (Dinev & Hart, 2006; Jiang, Heng, & Choi, 2013; Xu, Luo, Carroll, & Rosson, 2011; N. Lee & Kwon, 2015). Importantly, the privacy calculus framework accommodates the possibility that individuals may continue to express concerns about privacy while nonetheless choosing to disclose data when perceived benefits outweigh expected costs. In this perspective, sharing information does not automatically equate to a loss of privacy, nor does complete concealment guarantee optimal outcomes. Studies in online and e-commerce environments illustrate that consumers may pay premiums to transact with privacy-protective merchants, display growing concerns over time, adopt privacy-seeking behaviors, or resort to obfuscation strategies while still participating actively in digital markets (Boyd & Marwick, 2011; Goldfarb & Tucker, 2012; Kang, Brown, & Kiesler, 2013; Rainie et al., 2013; Stutzman, Gross, & Acquisti, 2013; Tsai, Egelman, Cranor, & Acquisti, 2011; Jentzsch, Preibusch, & Harasser, 2012; Preibusch, Kübler, & Beresford, 2013).

1.2.3 Uncertainty, bounded rationality, and limits of the privacy calculus

The privacy calculus literature typically assumes decision makers who are informed about the relevant risks and can form (possibly subjective) probabilities over privacy outcomes. In practice, however, privacy outcomes are often uncertain in a stronger sense: individuals may be unaware that data are being collected at all, may not know which entities have access to their information, and may have little understanding of how data will be processed or reused (Wakefield, 2013). This context is closer to Knightian uncertainty (Knight, 1921) than to risk: probabilities of breaches, secondary uses, or identity theft are neither known nor easily knowable.

These informational frictions open the door to systematic deviations from the predictions of a fully informed privacy calculus. A growing body of work documents the role of heuristics, framing effects, present bias, and other cognitive limitations in privacy-related decisions (Acquisti, Grossklags, & et al., 2007; Knijnenburg & Kobsa, 2013; Friik & Mittone, 2019; Acquisti, 2014). Bounded rationality (Acquisti & Grossklags, 2005) and limited attention may further lead individuals to underweight low-probability but high-impact privacy risks, or to focus on salient immediate benefits instead. From this perspective, observed inconsistencies between concerns and behavior may reflect not only contextual specificity but also structural features of the decision environment: complexity, opacity, delayed consequences, and asymmetric information.

The experimental design deliberately abstracts from several of these sources of uncertainty. Rather than attempting to reproduce the full complexity of real-world data flows, I construct a setting in which

the probabilities of data leaks are clearly stated, stable over time, and realized immediately after each decision. This simplification is partly dictated by standard experimental requirements of transparency and non-deception, but it also serves a methodological purpose: it creates a benchmark environment that is relatively favourable to the privacy calculus view. If individuals are ever in a position to perform informed trade-offs between financial returns and privacy risks, it is arguably in such a context. Behaviour in this simplified setting therefore provides an upper bound on the weight that individuals place on privacy risks when informational frictions and ambiguity are minimised.

1.2.4 Data leaks, financial information, and economic trade-offs

A distinct strand of literature examines data breaches, identity theft, and the economics of information security. Empirical work shows that the probability distributions of breach incidents and the subsequent misuse of stolen information are highly uncertain and difficult to communicate to consumers (Mann, 2015; Solove, 2005). From the consumer standpoint, insufficient awareness and inadequate protection of personal data can be interpreted as “insecurity” in Solove’s taxonomy, capturing failures to prevent leaks and unauthorized access (Solove, 2005). Within an economic framework, however, such insecurity may itself be the outcome of a trade-off between privacy and other objectives.

Payment system models developed by (Roberds & Schreft, 2009) and (Kahn & Roberds, 2008) illustrate these trade-offs in financial contexts. In (Roberds & Schreft, 2009), privacy losses from identity theft are weighed against the benefits of broad access to credit and efficient payment networks. Similarly, (Kahn & Roberds, 2008) analyze the tension between the desire to avoid intrusive monitoring of individuals and the need to combat transaction fraud. In both cases, some degree of privacy risk may persist even as technology improves, because eliminating it completely would require costly monitoring or restrictions that themselves reduce welfare. These models highlight that privacy losses and security failures can arise as an equilibrium outcome of economic incentives, rather than solely from poor design or lack of awareness.

This study contributes to this literature by implementing data leaks involving sensitive financial information in an experimental setting with known probability distributions. Participants face repeated decisions in which one option offers higher expected financial returns at the cost of an explicit risk that their own income, assets, debts, or other financial indicators will be revealed to peers. This allows us to observe directly how individuals trade off financial performance against the risk of a financial data leak, and to compare these trade-offs with the qualitative predictions of the above models. In contrast to the complex and opaque environments typically studied in observational work on data breaches, this experiment isolates the impact of transparent leak probabilities and immediate leak realizations on behavior.

1.2.5 Social influences and individual heterogeneity in privacy decisions

An emerging body of research points to the importance of social influences and individual heterogeneity in privacy behavior. Social norms, peer actions, and the visibility of others' experiences (such as being hacked or suffering a data leak) can affect both perceived risks and acceptable levels of disclosure (Krasnova et al., 2009; Nissenbaum, 2004). Yet relatively few experimental studies have explicitly investigated how observing others' privacy losses shapes one's own willingness to take privacy risks.

The experimental design addresses this gap by distinguishing between personal and others' leaks. Participants not only experience leaks of their own financial data but also observe when leaks affect other participants. This feature enables us to evaluate whether social spillovers—captured by aggregate leak events in previous rounds—have a distinct effect on subsequent investment choices over and above individual experiences. I thus connect the economics of privacy with broader experimental literatures on social learning, externalities, and moral or social preferences, in which observing harm to others can modify willingness to take risks or impose externalities.

Finally, this study relates privacy behavior to individual characteristics such as risk and loss aversion, financial situation, digital literacy, and psychometric measures of privacy concerns and attitudes. Prior work suggests that factors such as trust, perceived control, and information sensitivity shape privacy decisions (Dinev & Hart, 2006; Xu et al., 2011; Mothersbaugh et al., 2012). At the same time, global measures of privacy concern have often shown limited predictive power for concrete behaviors, feeding the debate on the privacy paradox (Acquisti et al., 2016; Dienlin & Trepte, 2015). By combining standard elicitation tasks (for risk and loss aversion) with a detailed financial questionnaire, a digital knowledge assessment, and psychometric scales adapted from (Dienlin & Trepte, 2015), I can assess to what extent context-specific perceptions—such as the perceived riskiness and sensitivity of the financial data at stake—better explain observed choices than broad attitudinal measures.

1.2.6 Research questions and hypotheses

The experiment and main analyses were pre-registered on AsPredicted (ID #154386).² Guided by the literature reviewed above and by this pre-analysis plan, the empirical work is structured around the following research questions and hypotheses.

First, I ask to what extent individuals trade off financial returns against transparent risks of leaking their own financial data in a dynamic setting. The privacy calculus framework implies that higher expected financial returns should increase the propensity to select the option carrying a privacy risk, whereas higher probabilities of a data leak should decrease it. Accordingly, the first hypothesis is:

²See pre-analysis plan at <https://aspredicted.org/x5vf-jhkd.pdf>.

Hypothesis 1 (Financial return and leak risk) *The probability that a subject chooses the option involving a potential data leak increases in its expected return relative to the safe option and decreases in the probability of a data leak.*

Second, I investigate how experimental conditions that exogenously modify the relative attractiveness of the privacy-risky option affect behavior. The induced-value treatments systematically shift the expected return of the option carrying the data leak risk while keeping risk and leak probabilities comparable. This leads to the second hypothesis:

Hypothesis 2 (Induced values of privacy risk) *For a given leak probability and baseline spread, subjects in the positive induced-value treatment are more likely to choose the option involving a potential data leak than subjects in the no induced-value condition, while subjects in the negative induced-value treatment are less likely to do so.*

Third, I study how data leak events affect subsequent choices over time. The design allows us to distinguish leaks affecting a given subject from leaks affecting other participants. Economic models of information security suggest that the perceived cost of a leak depends on the magnitude of the incident (Solove, 2005; Roberds & Schreft, 2009; Kahn & Roberds, 2008). Building on the idea that experienced loss and salient social information can shape risk-taking and externality-imposing behavior, I formulate:

Hypothesis 3 (Personal leaks versus magnitude of leaks) *Experiencing a leak of one's own financial data reduces, at least temporarily, the probability of choosing the privacy-risky option in subsequent rounds. Observing leaks affecting other participants also reduces this probability, and may have a distinct or stronger effect than personal leaks.*

Finally, I analyze how individual characteristics shape privacy-related decisions in this environment. The literature highlights the importance of risk and loss aversion, financial situation, digital literacy, and privacy attitudes for disclosure behavior (Dinev & Hart, 2006; Xu et al., 2011; Dienlin & Trepte, 2015). The pre-registered analyses include the predictive power of such measures for behavior in the investment task. I thus state:

Hypothesis 4 (Individual heterogeneity) *Higher risk aversion, greater perceived sensitivity and riskiness of the financial data at stake, and stronger context-specific privacy concerns are associated with a lower propensity to choose the privacy-risky option, whereas broad, generic measures of privacy concern have limited predictive power for behavior in the experiment.*

These hypotheses structure the empirical analysis that follows, in terms of both overall trade-offs between financial returns and privacy risk and the dynamic, social, and heterogeneous dimensions of privacy-related decision-making.

Table 1.1: Experimental design at a glance

Element	Value
Participants	186 participants (laboratory sessions).
Main task	60 rounds. Each round: endowment of 100 ECUs; choose between Options A and B; choose the fraction of the endowment to invest (uninvested amount yields 0).
Payoffs	Each option has three states (optimistic/realistic/pessimistic) with probabilities 25%/50%/25%. Within a round, the two options have the same payoff standard deviation; they differ only in expected return and privacy risk. Payoffs drawn from 30 scenarios.
Return spread	Difference in expected return (percentage points) varies across rounds: 0–5 p.p. (low), 5–10 p.p. (medium), >10 p.p. (high).
Privacy risk	In each round, exactly one option carries a leak probability $p_{\text{leak}} \in \{0, 5, \dots, 40\}\%$; the other option has $p_{\text{leak}} = 0$. If a leak occurs, it is realized immediately after the monetary outcome.
Leak content & audience	Leak reveals to other participants in the session: pseudonym and age + two randomly drawn items from seven pre-elicited financial indicators (income, financial wealth, real estate wealth, indebtedness status, indebtedness level, saving capacity, rental situation).
Treatments	Between-subject: Baseline; +4 p.p. shift in expected return of the leak option; –4 p.p. shift.
Payment	One round randomly selected for payment; conversion rate 12.5 ECUs per €1.

1.3 Experimental design and procedures

1.3.1 Allocation decision task

The core of the experiment is an allocation decision task comprising 60 rounds. Table 1.1 summarizes the key experimental parameters. In each round, participants received an endowment of 100 experimental currency units (ECUs) and decided how much of this endowment to allocate to one of two investment options. In the written instructions, both options were explicitly described as *risky investment opportunities* (Options A and B), and the task as choosing an investment and the amount to invest.³ Following Gneezy and Potters (1997), participants chose the fraction of the endowment to invest in the selected option, while any non-invested amount yielded no return. This feature limits the scope for trivial “opt-out” behavior driven purely by loss aversion.

Each option consisted of three possible states of nature—optimistic, realistic, and pessimistic—with fixed probabilities of 25%, 50%, and 25%, respectively. Each state yielded a monetary payoff, so that expected returns differed across options. The decision interface, which was presented as an investment screen, displayed for each option the three payoffs and their probabilities, the expected payoff (average return), and the difference in expected payoff between the two options, expressed in percentage points, exactly as described in the instructions (1.6.1). The two options within a round always shared the same standard deviation, ensuring comparable financial risk profiles. The only differences between the options were their expected returns and the presence of a privacy risk: in each round, one option was associated

³See the full instructions in Table 1.A.4.

with a strictly positive probability that the participant’s financial data would be leaked, while the other option was free of such risk.

The payoffs for both options were drawn from a set of 30 distinct scenarios. These scenarios were constructed so that, across the 60 rounds, the option involving a potential data leak financially dominated the safe option in exactly half of the cases and was dominated in the other half. This symmetry ensures that participants faced both financially attractive and unattractive privacy-risky options in equal proportions.

Throughout the task, I varied the spread in expected returns between the two options. This spread serves as a measure of the financial compensation offered for accepting the privacy risk and can be interpreted as a proxy for participants’ willingness to accept potential externalities induced by data leaks. The 30 scenarios were designed to span three levels of spread, expressed in percentage points: 10 scenarios with a low spread (between 0 and 5 p.p.), 10 with a medium spread (between 5 and 10 p.p.), and 10 with a high spread (above 10 p.p.).

The order of scenario presentation was randomly determined at the session level and kept identical for all participants within a session, as were the associated leak probabilities (see below). Thus, all participants within a session faced the same sequence of financial and privacy-risk profiles. After each round, participants received immediate feedback on the realized state and the resulting payoff. At the end of the experiment, one round was randomly selected for payment, with a conversion rate of 12.5 ECUs per euro.

1.3.2 Experimental treatments

Participants were randomly assigned to one of three between-subjects treatments summarized in Table 1.2. In all treatments, participants first completed a socio-demographic and financial questionnaire, as well as two incentivized tasks (Section 1.3.4). The treatments differed only in the induced value attached to the option carrying the privacy risk, in line with the induced-values framework of (V. L. Smith, 1976). This variation modifies financial incentives while preserving the structure of privacy risks.

In the baseline treatment, expected returns were determined solely by the 30 payoff scenarios described above. Treatment 1 added a systematic positive shift of 4 percentage points to the expected return of the option involving a potential data leak, whereas Treatment 2 applied a systematic negative shift of 4 percentage points to that option. This manipulation creates exogenous variation in the financial attractiveness of the privacy-risky option and allows us to study how stronger or weaker financial compensation affects the willingness to accept privacy risk.

To maintain comparability across treatments, I imposed a simple rule on expected return differences when the privacy-risky option financially dominated the safe option. In the baseline treatment, the

average difference in expected returns in such cases was 8 *p.p.*; in Treatment 1 and Treatment 2, the corresponding averages were 12 *p.p.* and 4 *p.p.*, respectively.

Table 1.2: Treatment overview

Condition	Induced value for the data-leak option
Baseline	None
Treatment 1	Systematic and positive
Treatment 2	Systematic and negative

Notes. This table reports the between-subjects treatment conditions.

1.3.3 Implementation of data leaks

Implementing data leaks in a laboratory environment raises both ethical and methodological challenges. In the design, the probability of a data leak is framed as an attribute of the investment options: for each decision interface, one of the two risky investments carries a strictly positive probability that some of the investor’s financial data will be disclosed, while the other investment is leak-free. In line with standard practices that avoid deception in economic experiments (Friedman & Sunder, 1994), all features of the data leaks were fully disclosed to participants before the start of the study. The primary objective was to examine how transparent risk probabilities, combined with financial incentives, shape privacy-related decisions. I therefore deliberately prioritized perfect information over real-world ambiguity.

In the original design, both the volume and the content of leaked data were to be randomly determined, in an attempt to mimic real-world uncertainty. To satisfy ethical constraints while preserving internal validity, I adopted a “degraded” leak implementation in the final design. Leaks were restricted to a subset of the data collected in the initial financial questionnaire, thereby reducing their potential impact while maintaining the essential structure of the privacy risk.

The final protocol was approved by the Institutional Review Board⁴ (IRB) associated with the Laboratoire d’Économie Expérimentale de Paris⁵ (LEEP). In the implemented design, a leak consisted of revealing to other participants the subject’s pseudonym, age, and two data items drawn at random from a pool of seven financial indicators: declared annual income, financial wealth, real estate wealth, indebtedness status, level of indebtedness, saving capacity, and rental situation. Focusing on purely economic variables follows (Mothersbaugh et al., 2012) and (Carrascal et al., 2013), who show that information sensitivity matters for disclosure decisions and that financial and offline data tend to be perceived as more sensitive and more valuable than many online traces.

⁴<https://www.parisschoolofeconomics.eu/fr/connaitre-pse/engagements-ethiques/integrite-scientifique/irb-institutional-review-board/>

⁵<https://leep.univ-paris1.fr/accueil.htm>

The probabilities of data leaks were determined at the session level by randomly drawing 30 values and associating them with the 30 payoff scenarios, before generating symmetric counterparts as described in Section 1.3.1. The leak probability parameter, denoted p_{leak} , took values between 0% and 40%, in 5 percentage point increments. These values were then held fixed across participants within a session. In each round where $p_{leak} > 0$, choosing the option associated with the privacy risk implied that the participant faced the stated probability that a leak would occur immediately after the financial outcome was realized.

Collecting sensitive financial data raises additional concerns about social desirability and misreporting. Sensitive questions about income and wealth, in particular, are prone to both underreporting and overreporting, as documented in survey research (Moore & Welniak, 2000; Angel, Disslbacher, Humer, & Schnetzer, 2019; Bound, Brown, & Mathiowetz, 2001). To mitigate these issues, I implemented a two-step data collection procedure. The financial questionnaire used to elicit the variables that could later be leaked was first administered as a pre-registration questionnaire prior to the lab session. The same questions were then asked again at the beginning of the experiment. Discrepancies between the two sets of answers were used ex post as a control for potential misreporting in the econometric analysis.

1.3.4 Questionnaires and additional tasks

Before entering the allocation decision task, all participants completed a comprehensive financial questionnaire designed to mirror standard data collection practices in the financial industry, in line with the MiFID II directive.⁶ The questionnaire covered socio-demographic characteristics, income and wealth, indebtedness, rental status, savings behavior, and self-reported financial knowledge and experience. As noted above, these data served both as potential leak content and as controls in the analysis.

The questionnaire also included two incentivized tasks to elicit preferences over risk and losses. Risk attitudes were measured using the multiple price list design of (Holt & Laury, 2002) (Figure 1.A.2 in the Appendix), while loss aversion was measured using a simple lottery choice task adapted from (Gächter, Johnson, & Herrmann, 2022) (Figure 1.A.4 in the Appendix). Payoffs from these tasks were realized at the end of the experiment. Summary statistics for the elicited preference parameters are reported in 1.6.1.

At the end of the main task, participants completed two additional questionnaires. The first was a digital knowledge test based on Europass materials,⁷ which I use as a proxy for digital literacy in the analysis. The second was a psychometric questionnaire adapted from (Dienlin & Trepte, 2015), extended with additional items tailored to the experimental context. This instrument captures multiple dimensions

⁶Directive 2014/65/EU of the European Parliament and of the Council of 15 May 2014 on markets in financial instruments.

⁷<https://europa.eu/Europass/fr>

of privacy, including informational, social, and psychological aspects, as well as participants' perceived risk of data leaks and self-reported privacy-related behaviors. Detailed wording, response scales, and aggregation procedures for these items are provided in 1.6.1.

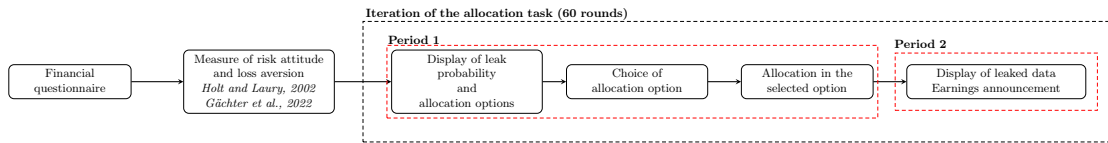
1.3.5 Recruitment and data collection procedures

The experiment was conducted at the LEEP, the experimental economics laboratory of Université Paris 1 Panthéon-Sorbonne and the Paris School of Economics. Participants were recruited from the laboratory's pool of volunteers. To limit age-related heterogeneity, individuals older than 50 were not invited. Sessions took place in December 2023; experimental data were collected between 8 and 22 December 2023.

Upon arrival, participants received written instructions and accessed the experimental software, which was implemented in oTree (D. L. Chen, Schonger, & Wickens, 2016), using personal identifiers created at the pre-registration stage. After completing the pre-registration financial questionnaire and the two incentivized preference tasks, participants were reminded of the nature of the data leaks and informed that participation was voluntary and could be discontinued at any time. None of the 186 enrolled participants chose to withdraw, and all provided explicit informed consent.

The main allocation task then started, followed by the digital knowledge and psychometric questionnaires. Throughout, I implemented strict data handling procedures to ensure that experimental data remained pseudonymous and that only approved variables could appear in potential leaks. The experiment was conducted in French and lasted on average 90 minutes. Participants received an average payment of €16.09 (SD = 3.77), approximately US\$17.78 at the exchange rate prevailing at the end of December 2023.⁸ Figure 1.1 summarizes the sequence of tasks and the timing of information and outcomes.

Figure 1.1: Flowchart of the experiment



1.4 Results

1.4.1 Sample descriptives

A total of 186 participants took part in the experiment. On average, they were 28.19 years old (SD = 7.63), and 51.07% were male. In terms of annual income, 56.45% reported earning between 0 and 10,000 euros, 16.12% between 10,000 and 25,000 euros, 22.58% between 25,000 and 45,000 euros, 3.22% between

⁸At the end of December 2023, the EUR–USD exchange rate was approximately 1 USD = 0.9 EUR.

45,000 and 60,000 euros, and 1.7% above 60,000 euros. Regarding education, 44.62% held a Master’s degree, 36.55% an undergraduate degree or BTEC Higher National Diploma, and 16.12% a secondary degree. Students constituted the majority of the sample (56.99%), followed by employees (16.66%), white-collar and highly qualified workers (10.75%), intermediate professions (6.45%), other occupations (5.37%), and craftsmen, shopkeepers, and business owners (3.76%).

To assess the quality of the financial and socio-demographic data, I compare answers given in the pre-registration questionnaire and in the in-lab questionnaire. For variables not eligible for leaks (age, gender, marital status, number of children, socio-professional category, education level, and postal code), the average consistency ratio is 96.39% (SD = 8.47). For variables that can be included in data leaks (see Section 1.3.3), the average consistency ratio is 82.53% (SD = 15.11); the difference in the distribution of the two ratios is statistically significant ($p < 0.001$). Allowing for a 2.5% deviation above and below, the specific consistency ratio is 83.33% (SD = 14.85). On average, participants spent 63.80 minutes on the main allocation task across sessions. Additional summary statistics and details on completion times are reported in the Appendix.

In the remainder of this section, I first study participants’ allocation decisions in the investment task, focusing on the trade-off between expected returns and the risk of a data leak. I then examine the role of leak realizations over time and the influence of individual characteristics, including psychometric measures of privacy attitudes and self-reported perceptions of risk.

1.4.2 Main trade-off between financial returns and privacy risk

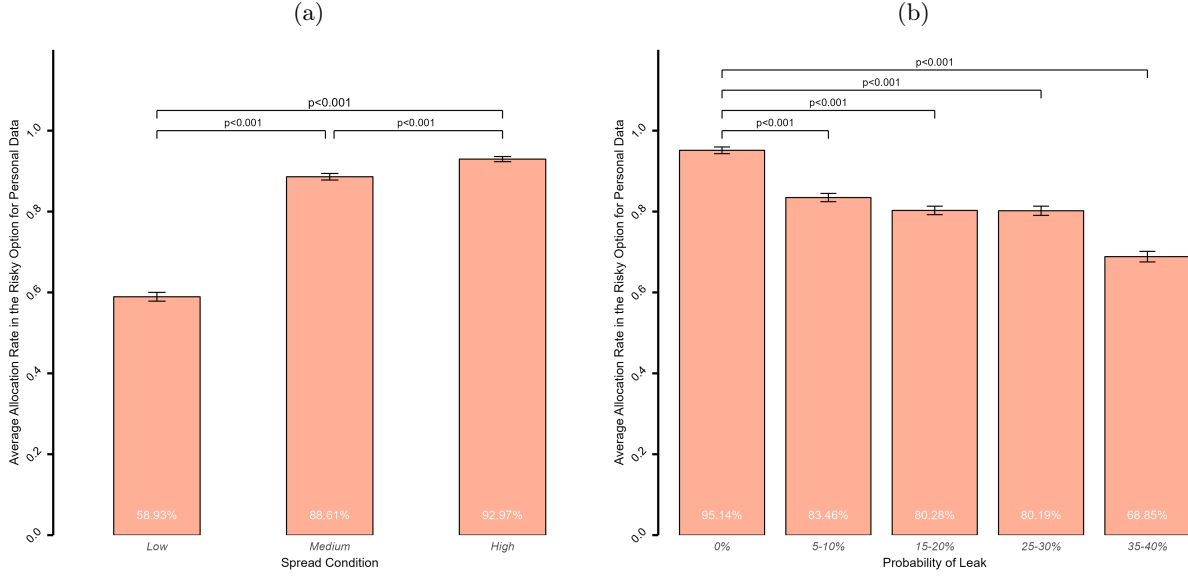
Descriptive evidence: spreads and leak probabilities

I begin by providing descriptive evidence on how participants trade off expected financial returns against the probability of a data leak in the investment task. In this subsection, I focus on rounds in which the option carrying the data leak risk financially dominates the safe option in terms of expected returns, so that choosing the safe option entails a (pure) privacy premium.

Result 1 (Trade-off between expected returns and leak risk). *Participants’ allocation decisions respond systematically to both the spread in expected returns and the probability of a data leak: they allocate more to the privacy-risky option when it offers a higher expected return, and allocate less to it as the leak probability increases.*

Support: Panel (a) of Figure 1.2 shows, for rounds in which the “leak” option offers a higher expected return than the safe option and $p_{leak} > 0$, the fraction of rounds in which participants allocate a positive amount of their endowment to the leak option. The underlying estimates come from a logistic regression

Figure 1.2: Average allocation choice in the risky option for personal data

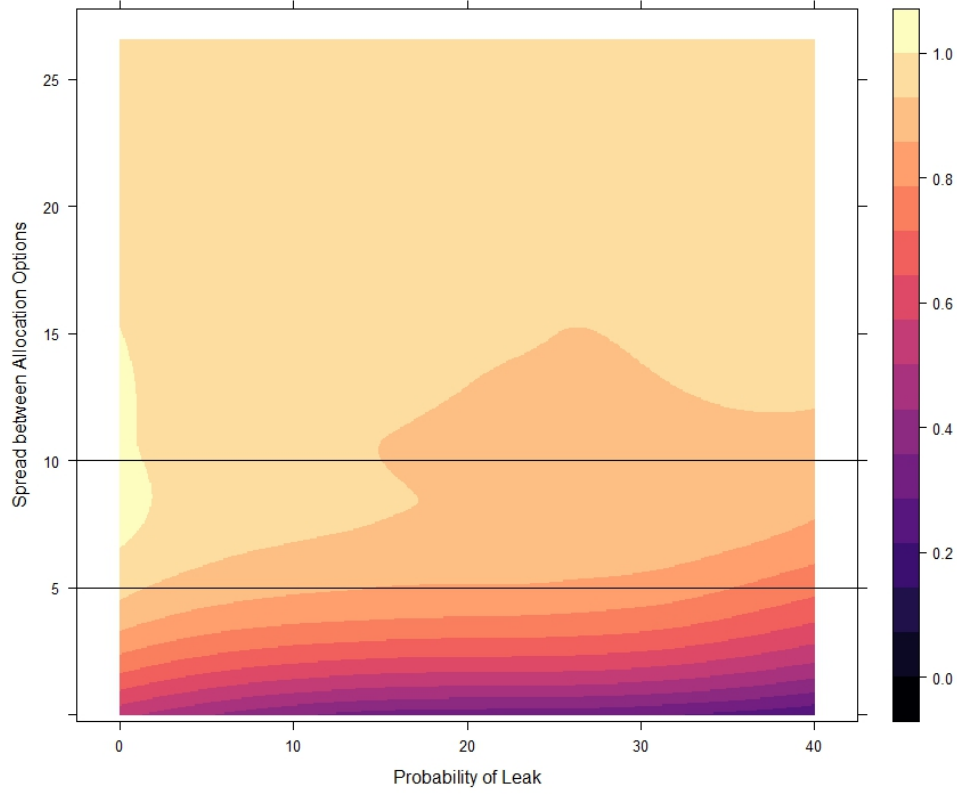


Notes. Panel (a) depicts the average allocation rate to the option carrying the data leak risk by spread category, conditional on this option dominating the safe one in terms of expected returns. Panel (b) shows the average allocation rate to the same option by probability level of the data leak risk. Rounds with $p_{leak} = 0$ are excluded in Panel (a) and included in Panel (b). Error bars indicate standard errors of the mean (SEM). p -values are based on logistic regressions of the binary allocation choice on spread (Panel a; $n = 5231$) and leak-probability categories (Panel b; $n = 5910$), with standard errors clustered at the participant level.

of the binary allocation choice on indicators for the three spread categories ($n = 5231$; robust standard errors clustered at the participant level). When the spread in expected returns between the leak and the safe option is low (0–5 percentage points), the average allocation rate to the leak option is 58.93%. This rate increases sharply with the spread: it reaches 88.61% in the Medium spread condition (AME = 0.221, $p < 0.001$) and 92.97% in the High spread condition (AME = 0.270, $p < 0.001$). The difference between the Medium and High spread conditions is also statistically significant (AME = 0.072, $p < 0.001$). In the investment frame of the task, participants are thus much more willing to accept the possibility of a data leak when the associated option offers a sufficiently higher expected monetary return.

Panel (b) of Figure 1.2 documents how the leak probability affects choices in rounds where the leak option dominates the safe one. Here I group rounds by the level of p_{leak} and report the fraction of allocations to the leak option within each group, based on a logistic regression of the binary allocation choice on leak-probability categories ($n = 5910$; clustered standard errors). When $p_{leak} = 0$, i.e. when no privacy risk is present, 95.14% of choices go to the option with the highest expected return. As the probability of a data leak increases, the allocation rate to the leak option declines monotonically: to 83.46% for $p_{leak} \in \{5, 10\}\%$ (AME = -0.249 , $p < 0.001$), 80.28% for $p_{leak} \in \{15, 20\}\%$ (AME = -0.290 , $p < 0.001$), 80.19% for $p_{leak} \in \{25, 30\}\%$ (AME = -0.299 , $p < 0.001$), and 68.85% for $p_{leak} \in \{35, 40\}\%$ (AME = -0.431 , $p < 0.001$). Even in this transparent environment, higher leak probabilities therefore discourage investment in the privacy-risky option.

Figure 1.3: Heatmap of average allocation to the risky option for personal data

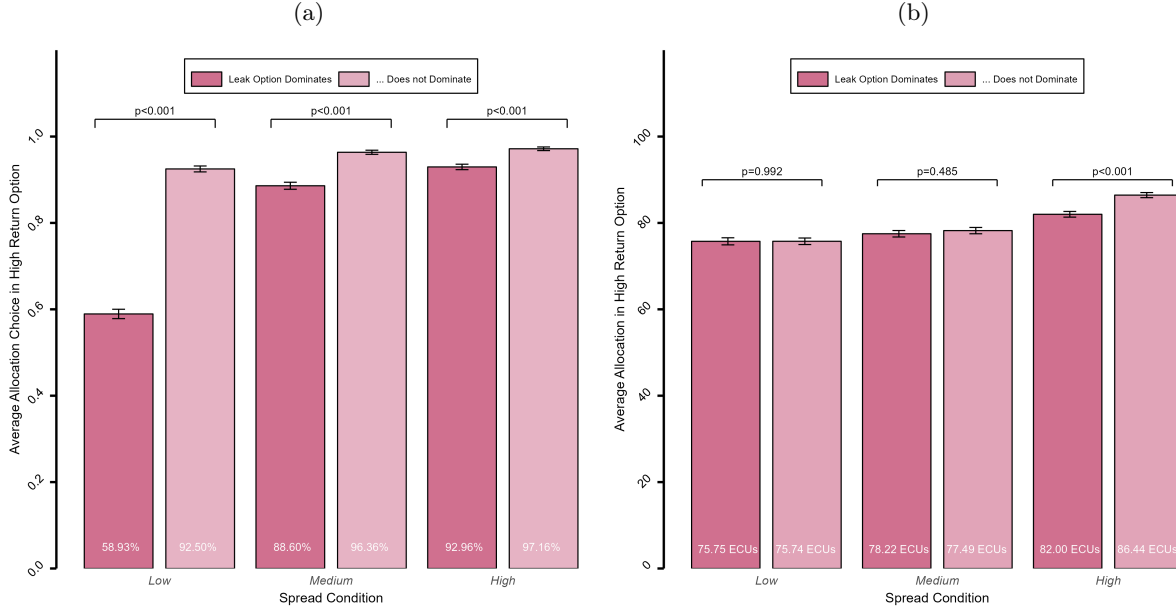


Notes. The figure illustrates the average allocation rate to the option carrying the data leak risk across combinations of leak-probability levels and spread categories, conditional on this option dominating the safe one in terms of expected returns ($n = 5910$). Horizontal lines indicate the distinct spread parametrizations introduced in Section 1.3.1.

Figure 1.3 combines spreads and leak probabilities and confirms that both dimensions matter jointly. When the spread is below 5 percentage points, participants tend to allocate more often to the safe option, regardless of the probability of a leak. For intermediate spreads (between 5 and 10 percentage points), there is a marked shift toward the safe option when p_{leak} exceeds 25%. At higher spreads (above roughly 15 percentage points), average allocation rates to the leak option become more uniform and generally high, suggesting that, beyond a certain return premium, potential gains from the leak option largely outweigh concerns about privacy risk for most participants.

Finally, Figure 1.4 contrasts choices across dominance relations. Panel (a) shows that the average allocation rate to the safe option when it offers the higher expected return is systematically higher than the average allocation rate to the leak option when *it* offers the higher expected return, within each spread category. In other words, participants react more strongly to financial dominance when it is aligned with privacy protection than when it requires accepting a data leak risk. Panel (b) shows that, conditional on choosing a given option, the average invested amount is broadly similar across scenarios, except in the High spread condition, where participants allocate significantly more when the safe option dominates

Figure 1.4: Average allocation choice and allocation level



Notes. Panel (a) compares the average allocation rate to the leak option when it dominates the safe option in terms of expected returns ($N = 5231$) with the average allocation rate to the safe option when it dominates the leak option ($N = 4571$), by spread category. Panel (b) compares the corresponding average allocation levels (conditional on investing). Error bars indicate standard errors of the mean (SEM). p -values are based on two-sided Mann-Whitney U tests.

than when the leak option dominates ($t(3009.6) = -5.0106$, $p < 0.001$). Taken together, these patterns suggest that participants take the privacy dimension into account when deciding whether to invest in the leak option, but that, once this discrete choice is made, the intensive margin of how much to invest is less sensitive to the presence of privacy risk.

Overall, the descriptive evidence in this subsection supports a systematic trade-off between expected financial returns and the probability of a data leak in the investment task. Participants are more likely to choose the privacy-risky option when the return premium is large and less likely to do so as the leak probability rises, providing a first, reduced-form validation of Hypothesis 1.

Econometric evidence: spread, leak probability, and treatments

Result 2 (Financial performance versus privacy protection). *When the option carrying the data leak risk weakly dominates the safe option in terms of expected return, participants' allocation decisions respond to both the leak probability and the return spread, but are substantially more sensitive to the latter, indicating that financial performance tends to outweigh privacy concerns.*

Support: Table 1.3 reports logit estimates of the probability of allocating to the option carrying the data leak risk, separately by treatment (Models 1–3) and for the pooled sample (Models 4–5). The

Table 1.3: Determinants of Allocation Choices: Experimental Parameters

	<i>Dependent variable: Data Leak Allocation</i>				
	Baseline (1)	Treatment 1 (+) (2)	Treatment 2 (-) (3)	Pooled (4)	Pooled (5)
Spread (Std.)	0.856*** (0.221)	1.776*** (0.309)	1.400*** (0.242)	1.481*** (0.178)	1.577*** (0.177)
Probability (Std.)	-0.442*** (0.079)	-0.225*** (0.065)	-0.119* (0.056)	-0.219*** (0.041)	-0.236*** (0.043)
Treatment 1 (+)				-0.740** (0.260)	-0.781** (0.265)
Treatment 2 (-)				-0.042 (0.256)	-0.183 (0.266)
Constant	1.857*** (0.293)	1.469*** (0.266)	1.985*** (0.299)	2.100*** (0.268)	2.838 (1.938)
Controls	No	No	No	No	Yes
Num. obs.	1,803	2,143	1,285	5,231	5,231
Pseudo- R^2 (McFadden)	0.115	0.299	0.122	0.192	0.237
Log Likelihood	-736.40	-809.51	-641.89	-2,210.39	-2,087.09

Notes. This table reports estimates from logistic regressions of the binary choice of whether to allocate the investment to the option carrying the data leak risk on a set of experimental parameters, conditional on the treatment (Models 1–3) and pooled across all treatments (Models 4–5). The sample is restricted to rounds with strictly positive data leak probability and non-negative spreads. Standard errors clustered at the participant level are reported in parentheses. “Probability (Std.)” and “Spread (Std.)” denote standardized values of, respectively, the probability of a data leak associated with the risky allocation option and the difference in expected returns (in percentage points) between the risky and the safe option, in rounds where the risky option weakly dominates in terms of expected return. “Controls” include age, gender, number of children, marital situation, maximum education level, professional category, financial literacy (PCA index), and response precision measures.

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

estimation sample is restricted to rounds in which (i) the leak probability is strictly positive and (ii) the option carrying the data leak risk weakly dominates the safe option in terms of expected return (non-negative spread). The first restriction ensures that a privacy trade-off is always present, so that the coefficient on leak probability can be interpreted as capturing how changes in an existing leak risk are associated with changes in the log-odds of choosing the risky option, rather than the discrete jump from no risk to some risk. The second restriction excludes rounds in which the leak option is strictly dominated in expected return, where almost all participants allocate to the safe option by construction, and variation in choices is largely mechanical.

Across all specifications, the standardized spread variable enters with a positive and highly significant coefficient, while the standardized leak probability enters with a negative and significant coefficient. In the baseline condition (Model 1), a one-standard deviation increase in the spread is associated with an increase of 0.856 in the log-odds of choosing the leak option, whereas a one-standard deviation increase in the leak probability reduces the log-odds by 0.442. In the pooled specification with controls (Model 5), the coefficient on spread rises to 1.577, compared with -0.236 for the leak probability. In terms of odds

ratios, this implies that a one-standard deviation increase in the spread multiplies the odds of choosing the leak option by roughly 4.8, while a one-standard deviation increase in the leak probability reduces those odds by about 21%.

The same qualitative pattern holds within each treatment arm. In the positive induced-value treatment (Model 2), the coefficient on spread (1.776) is more than seven times larger in absolute value than the coefficient on the leak probability (-0.225), and in the negative induced-value treatment (Model 3), the spread coefficient (1.400) likewise dominates the probability coefficient (-0.119). Adding socio-demographic and literacy controls in the pooled specification (Model 5) leaves the magnitudes of both coefficients essentially unchanged, suggesting that neither composition effects nor observable individual characteristics are driving these relationships. These results show that, within the domain where the leak-bearing option offers at least as high an expected return as the safe option, participants do react to higher leak probabilities by allocating less often to the risky option, but their behavior is much more elastic with respect to financial incentives. This pattern is consistent with the privacy calculus framework: participants tend to accept that risk when the expected monetary gain is sufficiently high, indicating that the marginal value they place on privacy protection is often outweighed by return considerations.

Turning to the induced-value treatments, the coefficients on the treatment dummies in Columns (4)–(5) do not support Hypothesis 2. Relative to the baseline condition, participants in the positive induced-value treatment are in fact less likely to choose the leak option (Treatment 1 coefficients around -0.75 , significant at the 1% level), while participants in the negative induced-value treatment do not differ significantly from the baseline once I condition on spread and leak probability. This pattern suggests that, in this setting, the induced-value manipulation does not generate an additional shift in privacy-risk taking beyond that already captured by the experimental parameters themselves.

1.4.3 Dynamic effects of data leaks

Result 3 (Dynamic effects of data leaks). *Past data leaks affect subsequent allocation decisions primarily through their aggregate, rather than individual, consequences: the number of other participants affected by a leak in the previous round significantly reduces the likelihood of investing in the privacy-risk option, whereas own past leaks have, at best, only modest and short-lived effects.*

Support: Table 1.4 reports fixed-effects logit estimates of the probability of allocating to the option carrying the data leak risk, pooling all treatments and controlling for experimental parameters (spread and leak probability), individual fixed effects, and round fixed effects (Models 1–5). The different specifications successively introduce indicators capturing the timing and persistence of individual leaks, as well as the number of leaks experienced by other participants in the previous round.

I first consider the impact of participants’ own data leaks. The variable *First Leak* = 1 takes value 1 if

Table 1.4: Determinants of Allocation Choices: Privacy Outcomes

	<i>Dependent variable: Data Leak Allocation</i>				
	(1)	(2)	(3)	(4)	(5)
First Leak _{t-1}	0.066 (0.114)				
First Leak _{t-1} (Cont.)		-0.216 (0.157)			
Individual Leak _{t-1}			-0.072 (0.115)		0.030 (0.121)
Others Participants' Leaks _{t-1}				-0.040** (0.001)	-0.041** (0.001)
Constant	-1.726*** (0.313)	-1.702*** (0.312)	-1.724*** (0.313)	-1.631*** (0.316)	-1.630*** (0.316)
Experimental Parameters	✓	✓	✓	✓	✓
FE Individuals & Rounds	✓	✓	✓	✓	✓
Num. obs.	10974	10974	10974	10974	10974
PseudoR ² (McFadden)	0.592	0.592	0.592	0.593	0.593
Log Likelihood	-2987.2	-2985.9	-2987.0	-2983.4	-2983.4

Notes. This table reports estimates from fixed-effects logistic regressions of the binary choice of whether to allocate the investment to the option carrying the data leak risk on a set of privacy consequences derived from the data leak risk present in the experiment. Standard errors clustered at the participant level are reported in parentheses. Experimental parameters include "Probability" which refers to the probability level associated with the risky allocation option and "Spread" which indicates the difference in percentage points between the expected return of the risky and safe allocation option. Despite not being displayed in this table, experimental parameters are standardized and estimates are similar to the ones presented earlier. Both individual and round fixed effects are included.

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

a participant experienced their first data leak in the previous round and 0 otherwise, and *First Leak_{t-1}* (*Cont.*) remains equal to 1 in all subsequent rounds once a first leak has occurred. Across Models (1) and (2), neither of these variables is statistically significant, suggesting that neither the immediate local effect of a first leak nor its persistence over later rounds materially changes the probability of choosing the leak option once I condition on individual and round fixed effects and on the financial incentives. Likewise, the indicator *Individual Leak_{t-1}*, which captures any own leak in the previous round (irrespective of whether it is the first), is not significantly associated with subsequent leak allocations in Models (3) and (5).

By contrast, the variable *Others Participants' Leaks_{t-1}*, defined as the number of other participants in the same session who experienced a data leak in the previous round, is both negative and statistically significant coefficient in Models (4) and (5). This implies that, holding constant the experimental parameters and unobserved individual heterogeneity, an increase in the number of peers affected by a leak in the prior round is associated with a lower propensity to allocate to the option carrying the leak risk. In other words, participants appear to react more strongly to the magnitude of collective leaks than to their

own individual experience of a leak. Given the inclusion of rich controls and fixed effects, the magnitude of these coefficients is moderate, but the sign and statistical significance are robust, indicating a “crowd effect” whereby observing (or being informed about) others’ exposures makes participants temporarily more cautious in their privacy-related investment decisions.

Table 1.5: Determinants of Allocation Choices: Privacy Outcomes (Cont.)

	<i>Dependent variable: Data Leak Allocation</i>					
	(1)	(2)	(3)	(4)	(5)	(6)
First Leak _{<i>t</i>-1}	-0.045 (0.029)		-0.062* (0.030)		-0.088** (0.035)	
First Leak _{<i>t</i>-1} (Cont.)		-0.063** (0.021)		-0.086** (0.025)		-0.155*** (0.030)
Experimental Parameters	✓	✓	✓	✓	✓	✓
FE Individuals	✓	✓	✓	✓	✓	✓
Num. obs.	1245	1245	946	946	614	614
PseudoR ² (McFadden)	0.747	0.751	0.837	0.842	0.902	0.913
Log Likelihood	-254.1	-250.6	-163.0	-158.8	-97.60	-87.01

Notes. This table reports estimates from fixed-effects logistic regressions of the binary choice of whether to allocate the investment to the option carrying the data leak risk on a set of privacy consequences derived from the data leak risk present in the experiment and conditioned on the event window considered: 4 pre- and post-event rounds in Models 1-2, 3 pre- and post-event rounds in Models 3-4, and 2 pre- and post-event rounds in Models 5-6. Standard errors clustered at the participant level are reported in parentheses. Experimental parameters include "Probability", which refers to the probability level associated with the risky allocation option, and "Spread", which indicates the difference in percentage points between the expected return of the risky and safe allocation option. Despite not being displayed in this table, experimental parameters are standardized and estimates are similar to the ones presented earlier. Individual fixed effects are included.

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

To further explore the dynamic impact of individual leaks and to alleviate some of the identification challenges arising in the full panel, I complement this analysis with an event-study design centered around the first leak experienced by each participant. In this approach, the occurrence of the first leak is treated as an event, and I restrict attention to symmetric windows of rounds around that event. Table 1.5 reports fixed-effects logit estimates for windows of four pre- and post-event rounds (Models 1–2), three pre- and post-event rounds (Models 3–4), and two pre- and post-event rounds (Models 5–6). As before, all specifications control for the experimental parameters and individual fixed effects.

Within the widest event window (four rounds before and after the first leak), the local effect of the first leak, captured by *First Leak_t - 1*, is not statistically significant (Model 1), whereas the persistent indicator *First Leak_t - 1 (Cont.)* is negative and significant (Model 2), indicating that, once the first leak has occurred, participants become less likely to choose the leak option in subsequent rounds. When I narrow the event window to three pre- and post-event rounds (Models 3–4), both the local and persistent measures become statistically significant, with negative coefficients. For the shortest window (two pre- and post-event rounds; Models 5–6), the estimated coefficients on both the local and persistent indicators

are again negative and increase in absolute value, suggesting a stronger reduction in leak allocations in the immediate aftermath of the first leak. These patterns indicate that individual leaks do induce a behavioral adjustment away from the risky option in the short run, but that this effect attenuates as I move further away from the event and when the full panel (including rounds without leak risk or with dominated risky options) is considered.

Overall, the panel and event-study evidence point to a nuanced picture of dynamic privacy behavior. On the one hand, aggregate leak events—captured by the number of other participants affected in the previous round—consistently reduce willingness to accept privacy risk, even after controlling for incentives and fixed effects. On the other hand, the impact of own leaks is modest and short-lived, becoming clearly visible only when I zoom in on a narrow window around the first leak. Within the framework of privacy calculus, these findings suggest that while individuals do react to salient privacy shocks, especially when many others are simultaneously affected, these reactions are not strong enough to permanently overturn the dominant role of financial incentives documented in the previous result.

1.4.4 Individual heterogeneity

I complete the analysis by focusing on the influence of individual characteristics on the behaviors observed in the experiment and provide tentative evidence to verify whether the privacy paradox may arise due to significant informational asymmetry concerning privacy risks and outcomes by including in this analysis the psychometric measures from (Dienlin & Trepte, 2015). All the Likert scale measures used in the post-experimental survey are described in 1.6.1.

Economic and behavioral characteristics

Result 4 (Economic and behavioral characteristics). *Conditional on experimental parameters and basic demographics, observed economic and behavioral characteristics exhibit only limited explanatory power for allocation decisions: neither digital knowledge nor the proxies for loss and risk aversion significantly predict allocations to the data-leak option, while economic conditions matter only weakly, with a small and imprecise pattern suggesting that participants with significant property assets are more likely to accept leak risks.*

Support: Table 1.A.9 reports logit estimates of the probability of allocating to the option carrying the data leak risk on the restricted sample used in Section 1.4.2 (rounds with strictly positive leak probability and non-negative spreads), always controlling for the experimental parameters (spread and leak probability, standardized) and for the full set of socio-demographic and literacy controls.⁹ Columns

⁹As in previous tables, the coefficients on the experimental parameters are not reported but remain very similar in sign and magnitude to those presented earlier.

(1)–(4) add digital knowledge and behavioral proxies for loss and risk aversion; Columns (5)–(6) introduce pre-registered economic characteristics.

Across all specifications, the coefficient on *Digital Knowledge* is small, negative, and statistically insignificant. This suggests that, conditional on the financial incentives and basic demographics, participants' general digital skills do not translate into systematically different propensities to accept the data leak risk.

Turning to behavioral characteristics, Columns (2)–(4) and (6) introduce measures of *Loss Aversion* and *Risk Aversion*. The estimated coefficients on these variables are uniformly negative, indicating that more loss- or risk-averse individuals are, if anything, less likely to allocate to the leak option, but they are not statistically significant at conventional levels in any of the reported specifications. Adding the full set of economic variables in Column (6) does not appreciably change these estimates. Overall, I therefore do not find robust evidence that differences in these behavioral measures meaningfully shift the privacy–return trade-off, once I account for the structure of the task and observable demographics.

Columns (5) and (6) focus on economic characteristics that were pre-registered as potential determinants of privacy behavior. Using the lowest income group (0–10k euros) as the reference category, none of the higher income brackets exhibits a statistically significant association with the probability of selecting the leak option. The picture is similar for most economic variables. *Home Ownership*, *Rent*, *Financial Wealth*, *Savings*, and *Indebtedness* all have coefficients fail to reach statistical significance. The only economic characteristic that shows a consistently significant association is *Property Assets*. In both Columns (5) and (6), the coefficient on property assets is positive and statistically significant at the 5% level (0.263 and 0.274, respectively, with s.e. 0.129). This pattern suggests that participants who report owning positive property assets are somewhat more inclined to choose the leak option, conditional on the experimental parameters and other controls. However, the effect remains modest in size.

Taken together, these results indicate that, within the experimental framework, observable economic and behavioral heterogeneity plays only a secondary role in shaping privacy-related allocation decisions. The main drivers of behavior remain the experimental parameters themselves while differences in digital knowledge, loss and risk attitudes, and economic characteristics add little additional explanatory power. From the perspective of the privacy calculus, this suggests that, at least in this controlled setting with immediate privacy outcomes, most participants apply a broadly similar trade-off between monetary returns and privacy risks, irrespective of their economic circumstances or standard behavioral traits.

Psychometric and self-reported measures

Result 5 (Privacy concerns and self-perceptions). *Ex-post privacy concerns and self-reported attention to the risk of data leaks are negatively associated with choosing the leak option, whereas past*

Table 1.6: Determinants of Allocation Choices: Psychometric and self-reported measures

	<i>Dependent variable: Data Leak Allocation</i>		
	(1)	(2)	(3)
<i>Psychometrics measures</i>			
Privacy Concerns	-0.095 (0.050)		-0.113* (0.048)
<i>Self-reported measures</i>			
Information Accuracy		-0.463** (0.159)	-0.490** (0.152)
Data Leak Experienced		0.098 (0.144)	0.071 (0.138)
Interest in Others' Leaks		0.128 (0.142)	0.221 (0.139)
Perceived Risk		0.170 (0.091)	0.166 (0.090)
Leak Influence		-0.436* (0.170)	-0.387* (0.170)
Increased Awareness		-0.100 (0.159)	-0.161 (0.163)
Constant	-3.426 (4.368)	-0.635 (4.087)	-3.592 (4.170)
Controls	Yes	Yes	Yes
Num. obs.	5231	5231	5231
Log Likelihood	-1642.5	-1620.17	-1613.08
PseudoR ² (McFadden)	0.399	0.408	0.41

Notes. Logistic regressions of the choice to allocate to the data-leak option. The sample is restricted to rounds with strictly positive data-leak probability and non-negative spreads. Standard errors (in parentheses) are clustered at the participant level. Experimental parameters (probability and spread) are included but not reported. "Controls" include age, gender, number of children, marital situation, maximum education level, professional category, financial literacy (PCA index), and response precision measures.

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

personal exposure to data leaks and curiosity about others' leaks show no robust relationship with allocation behavior.

Support: Table 1.6 reports logit estimates of the probability of allocating to the option carrying the data leak risk on the same restricted sample as before (rounds with strictly positive leak probability and non-negative spreads), always controlling for the experimental parameters (standardized spread and leak probability) and for the full set of socio-demographic and literacy controls. All psychometric and self-reported variables are measured at the very end of the experiment, after participants have completed the allocation task. The coefficients should therefore be interpreted as correlational patterns between ex post attitudes/perceptions and earlier choices, rather than as clean causal effects.

The first column focuses on a composite *Privacy Concerns* index, constructed using principal component analysis on the ten items from (Dienlin & Trepte, 2015). In Model (1), the coefficient on this index is negative but only marginally significant (-0.095 , s.e. 0.050). When I add the full battery of self-reported perceptions in Model (3), the coefficient becomes slightly larger in magnitude and statistically significant

at the 5% level (-0.113 , s.e. 0.048). This pattern indicates that participants who score higher on the general privacy concerns factor are less likely to allocate to the leak option, even after conditioning on the financial incentives and individual controls. In other words, ex-post stated privacy concerns are aligned with more privacy-protective behavior within the allocation task.

Columns (2) and (3) then incorporate a set of self-reported measures elicited in the post-experiment questionnaire. *Information Accuracy* captures how precise participants report having been in their questionnaire responses (from “not at all” to “very precise”). Its coefficient is negative and statistically significant across Models (2) and (3) (e.g. -0.490 , s.e. 0.152 in Model (3)), suggesting that participants who claimed to have answered more accurately and carefully are less likely to select the leak option. Similarly, *Leak Influence*, which measures the self-reported extent to which the risk of a data leak affected investment decisions (from “not at all” to “very strongly”), is also negatively associated with choosing the leak option (-0.387 , s.e. 0.170 in Model (3)). These two variables can be viewed as ex post meta-cognitive assessments of how seriously participants took the privacy dimension during the task; their negative coefficients are consistent with the idea that those who report having integrated the risk of leaks into their reasoning are also those who, in practice, avoided the risky option more often.

By contrast, the remaining self-reported measures do not display robust, statistically significant associations with allocation decisions. Having previously experienced a personal data leak (*Data Leak Experienced*) is not significantly related to the propensity to choose the leak option (coefficient 0.071 , s.e. 0.138 in Model (3)), suggesting that prior real-world exposure, as captured by this coarse indicator, does not strongly shape behavior in the experimental environment. Self-reported *Interest in Others' Leaks* during the experiment (from “not at all” to “very strongly”) enters with a positive but non-significant coefficient (0.221 , s.e. 0.139).

The coefficient on *Perceived Risk*—how dangerous participants judge the risk of a data leak in this experiment—is positive but only marginally significant (0.166 , s.e. 0.090). Given the underlying scale, where higher values correspond to perceiving the leak as less dangerous (shifting from “very dangerous” toward “not dangerous”), this positive sign is consistent with the idea that participants who downplay the danger of data leaks are more likely to choose the leak option. Finally, *Increased Awareness*, which asks whether participants feel more aware of data-leak risks after the experiment (from “not at all” to “extremely”), has a negative but statistically insignificant coefficient, indicating no robust link between ex-post awareness and prior allocation behavior once the other variables are taken into account.

Taken together, the results provide mixed support for the pre-registered hypotheses. For Hypothesis 1 (financial return and leak risk), the evidence in Section 1.4.2 provides clear support: higher expected-return spreads strongly increase, and higher leak probabilities decrease, the likelihood of choosing the leak option. By contrast, Hypothesis 2 (induced values of privacy risk) is not supported: once I control

for spread and leak probability, subjects in the positive induced-value treatment are in fact less likely to choose the leak option than in the baseline, and those in the negative induced-value treatment do not differ significantly. For Hypothesis 3 (personal leaks versus magnitude of leaks), the evidence is mixed: as shown in Section 1.4.3, the number of other participants affected by a leak in the previous round robustly reduces subsequent investment in the privacy-risk option, whereas own leaks have at best modest and short-lived effects, which become visible only in narrow event windows around the first leak. Finally, Hypothesis 4 (individual heterogeneity) is only partially supported: as discussed in Section 1.4.4, I find little evidence that risk or loss aversion and most economic characteristics systematically shape allocation choices, while a generic privacy-concern index does exhibit a modest but statistically significant association with behavior, contrary to the last part of the hypothesis.

Beyond these hypothesis-by-hypothesis assessments, the contribution of the experiment lies less in “re-establishing” the privacy calculus and more in documenting how it operates in a specific, rich environment: a dynamic setting with instantaneous privacy outcomes, a persistent but ethically constrained leak risk, a sequence of repeated financial–privacy trade-offs, and visible leaks affecting other participants. Within this framework, I show that financial incentives remain the primary driver of behavior, but that they interact in nuanced ways with realized privacy shocks, information about others’ leaks, and ex-post privacy attitudes—a combination that has rarely been examined jointly in a single experimental design.

1.5 Conclusion

This paper has examined how individuals trade off financial performance against the protection of their own financial data in a dynamic environment with time-varying leak probabilities and instantaneous privacy outcomes. I implemented a repeated investment task in which one option could offer higher expected returns at the cost of an explicit risk that sensitive financial information would be leaked to peers, while the alternative option was privacy-safe. Within this framework, I combined rich experimental variation in expected returns and leak probabilities with realized leak events and detailed measures of individual characteristics and privacy attitudes.

The first main finding is that participants behave in a way that is consistent with a privacy–return trade-off, even when the data at stake are financial and leak probabilities are fully transparent. Higher spreads in expected returns markedly increase, and higher leak probabilities reduce, the likelihood of choosing the privacy-risky option. However, the estimated elasticities are strongly asymmetric: choices respond much more to changes in expected returns than to changes in leak probabilities. Moreover, the induced-value treatments designed to shift the financial attractiveness of the leak option do not generate the pattern predicted in the pre-analysis plan. If anything, the positive induced-value treatment

is associated with a lower propensity to choose the leak option once spread and probability are controlled for, and the negative induced-value treatment does not differ significantly from the baseline. Taken together, these results suggest that, in this benchmark setting with explicit probabilities and immediate consequences, financial incentives are the dominant force shaping behavior, and additional induced-value manipulations add little over and above the structure of payoffs and risks itself.

The second main finding concerns the dynamics of social spillovers. Using both panel regressions and an event-study design around the first personal leak, I document that one's own leaks have only modest and short-lived effects on subsequent choices. Participants temporarily reduce their propensity to choose the leak option after their first leak, but this adjustment attenuates as the window around the event widens and essentially disappears in the full panel once I control for individual and round fixed effects. By contrast, the magnitude of leaks affecting other participants in the previous round robustly predicts behavior: when more peers experience a leak, participants are less likely to choose the privacy-risky option in the next round. This "crowd effect" points to the importance of public signals and social information in privacy-related decision-making, over and above individual experience of harm.

Third, I find only limited evidence that standard measures of economic variables and behavioral traits systematically shape privacy choices in this environment. Risk and loss aversion, digital knowledge, income, and most economic characteristics are not robustly associated with the propensity to choose the leak option once experimental parameters and basic controls are taken into account. By contrast, a generic privacy-concern index (constructed from (Dienlin & Trepte, 2015)) and a small set of ex post self-reports—such as perceived response accuracy and the self-reported influence of leak risk on decisions—are modestly but consistently aligned with more privacy-protective behavior. These patterns speak to the privacy paradox debate: in this setting with instantaneous privacy outcomes and explicit probabilities, broad attitudinal measures are not completely uninformative, but their predictive power remains small relative to the direct effect of financial incentives and leak probabilities.

From a policy and design perspective, the experiment highlights several points. First, even when probabilities are clearly stated, and leaks involve sensitive financial information, many individuals are willing to accept non-trivial increases in leak risk for relatively moderate gains in expected return. This has implications for open banking and finance initiatives: if financial incentives are strong enough, transparency about leak probabilities alone may not suffice to induce conservative behavior. Second, the strong response to aggregate leak events suggests a potential role for public disclosure and breach-notification policies—not only as accountability tools, but also as mechanisms that can temporarily shift demand away from privacy-risky options following salient incidents. Third, the weak role of standard economic and behavioral heterogeneity, combined with the modest but coherent contribution of privacy attitudes, suggests that targeting interventions solely on the basis of observable characteristics may be difficult, and

that context-specific communication about risks may matter more than generic awareness campaigns.

Naturally, these conclusions come with limitations. The laboratory environment involves ethically constrained leaks to a small group of peers, not to the wider public or to firms, and the stakes, while salient, are modest compared to those associated with major real-world financial breaches. The sample is relatively young and educated, and the task abstracts from many forms of uncertainty and opacity present in actual data ecosystems. In that sense, the behavior observed here should be viewed as an upper bound on the extent to which individuals can engage in a transparent privacy calculus when probabilities are known and outcomes are immediate.

Future work could relax some of these constraints and explore complementary dimensions of privacy decision-making. One natural extension is to introduce ambiguity over leak probabilities or over the future uses of leaked data, to examine how ambiguity attitudes interact with financial incentives. Another is to vary the audience and consequences of leaks, for instance, leaks to firms, employers, or platforms rather than to anonymous peers, and to study how this affects the perceived cost of disclosure.

A further extension is to connect the social dimension highlighted by the data to the literature on punitiveness and norm enforcement. The finding that participants respond to others' leak realizations could reflect not only information updating, but also moralized reactions to norm violations, stigma, or a desire to sanction behaviors perceived as socially harmful. Building on evidence of costly punishment and social norm enforcement in economic environments (Fehr & Gächter, 2000; Herrmann, Thoni, & Gächter, 2008; Gächter & Herrmann, 2009; Boulu-Reshef & Monnier, 2024), future experiments could allow participants to impose costs on those who choose the privacy-risky option or who experience leaks, or could elicit approval/disapproval and reputational consequences. This would help distinguish social learning from punitive motives and clarify whether privacy-related choices are shaped by the threat of social sanctions.

Finally, combining dynamic leak environments with richer social structures and long-run outcomes could help bridge the gap between stylized experimental settings and the complex, evolving data ecosystems of the digital economy.

1.6 Appendix

1.6.1 Experimental Design

Decision Interface

Figure 1.A.1: Decision Interface

Allocation Decision [4/60]

Option A				Option B		
Risk of data leak	No risk			40%		
Scenarios	Unfavorable	Realistic	Favorable	Unfavorable	Realistic	Favorable
Probability	(25%)	(50%)	(25%)	(25%)	(50%)	(25%)
Yield	-4%	12%	32%	-8%	8%	28%
Expected Return	13%			9%		
Spread	4 percentage points					
Allocation decision	☐ A ☐ B					

Allocation amount

You have an endowment of **100,00** ECUs.

Specify the amount you wish to allocate in the option chosen above.

Additional Tasks

Figure 1.A.2: Risk Aversion Elicitation Task à la Holt & Laury (2002)

Initial Task No. 1

In this task, for each row you must choose between two lotteries, A and B, that are offered to you. Lottery A is characterized by a reward of 2 euros or 1.60 euros along with the associated probabilities for each reward, and lottery B is characterized by a reward of 3.85 euros or 0.10 euros along with the associated probabilities for each reward.

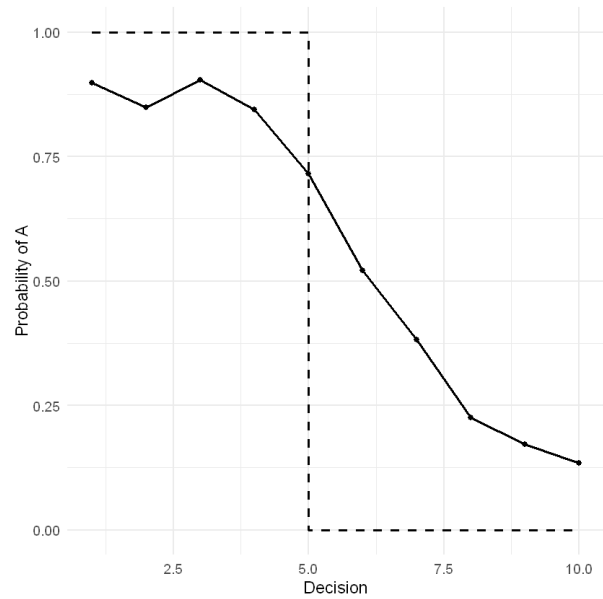
You have 10 choices to make (between A and B), one for each row. The amounts for the different lotteries will always be the same, but the probabilities will change. The lower you go in the table, the higher the probability associated with the larger reward.

Your reward will be determined by the random selection of one of the rows, and, depending on your choice, by the outcome of the lottery you selected.

#	A				B			
1	1/10th of 2,0€	or	9/10th of 1,6€	☐ A ☐ B	1/10th of 3,85€	or	9/10th of 0,1€	
2	2/10th of 2,0€	or	8/10th of 1,6€	☐ A ☐ B	2/10th of 3,85€	or	8/10th of 0,1€	
3	3/10th of 2,0€	or	7/10th of 1,6€	☐ A ☐ B	3/10th of 3,85€	or	7/10th of 0,1€	
4	4/10th of 2,0€	or	6/10th of 1,6€	☐ A ☐ B	4/10th of 3,85€	or	6/10th of 0,1€	
5	5/10th of 2,0€	or	5/10th of 1,6€	☐ A ☐ B	5/10th of 3,85€	or	5/10th of 0,1€	
6	6/10th of 2,0€	or	4/10th of 1,6€	☐ A ☐ B	6/10th of 3,85€	or	4/10th of 0,1€	
7	7/10th of 2,0€	or	3/10th of 1,6€	☐ A ☐ B	7/10th of 3,85€	or	3/10th of 0,1€	
8	8/10th of 2,0€	or	2/10th of 1,6€	☐ A ☐ B	8/10th of 3,85€	or	2/10th of 0,1€	
9	9/10th of 2,0€	or	1/10th of 1,6€	☐ A ☐ B	9/10th of 3,85€	or	1/10th of 0,1€	
10	10/10th of 2,0€	or	0/10th of 1,6€	☐ A ☐ B	10/10th of 3,85€	or	0/10th of 0,1€	

Continue

Figure 1.A.3: Proportion of safe choices in each decision (Holt & Laury (2002))



Notes. Data represent the average choice probabilities (solid line with dots) alongside the risk-neutral prediction (dashed line). 40 of 186 participants (21.5%) exhibited internally inconsistent choice patterns—an inconsistency rate comparable to that commonly reported in standard risk-elicitation protocols.

Figure 1.A.4: Loss Aversion Elicitation Task à la Gächter et al. (2022)

Initial Task No. 2

In the following table, you will find a list of coin toss games with different rewards. **The rewards differ depending on the amount you lose if the coin lands on heads or tails.** For each row, you must indicate whether you refuse or accept to toss the coin.

To determine your payment, one of the six rows will be selected at random. If you indicated that you wanted to toss the coin for the randomly selected row, the coin toss will be performed and will determine your winnings or losses.

#		
1	If the coin lands on heads, you lose 2€ ; if it lands on tails, you win 6€ .	☐ Refuser ☐ Acceptor
2	If the coin lands on heads, you lose 3€ ; if it lands on tails, you win 6€ .	☐ Refuser ☐ Acceptor
3	If the coin lands on heads, you lose 4€ ; if it lands on tails, you win 6€ .	☐ Refuser ☐ Acceptor
4	If the coin lands on heads, you lose 5€ ; if it lands on tails, you win 6€ .	☐ Refuser ☐ Acceptor
5	If the coin lands on heads, you lose 6€ ; if it lands on tails, you win 6€ .	☐ Refuser ☐ Acceptor
6	If the coin lands on heads, you lose 7€ ; if it lands on tails, you win 6€ .	☐ Refuser ☐ Acceptor

Continue

Table 1.A.1: Acceptance Behavior by Lottery Choice Category

Lottery Choice Category	N	%
Reject All Lotteries	7	3.8
Accept #1; Reject #2–#6	24	12.9
Accept #1–#2; Reject #3–#6	63	33.9
Accept #1–#3; Reject #4–#6	51	27.4
Accept #1–#4; Reject #5–#6	19	10.2
Accept #1–#5; Reject #6	9	4.8
Accept All Lotteries	13	7.0
Inconsistent Choice Patterns	22	11.8

Questionnaires

Table 1.A.2: Preliminary questionnaire

Question	Possible answer(s)
Socio-demographic questions	
What is your age?	<i>Numeric</i>
What is your gender?	Male Female Non-binary Prefer not to say
What is your marital status?	Single Cohabiting Married Divorced Widowed Prefer not to say
Do you have any dependent children?	None 1 child 2 children 3 or more children

	Prefer not to say
Which occupational category best describes you?	Farmer Artisan/merchant/business owner Executive/professional Intermediate profession Employee Worker Student Retired Other
What is your highest level of education completed?	Primary or no formal education Secondary Bachelor's degree or BTS Master's degree Doctorate
Economic Information	
What is the postal code of your tax residence?	<i>Numeric</i>
What is your net annual salary?	0–10 000 € 10,000–25,000€ 25,000–45,000€ 45,000–60,000€ 60,000–100,000€ Over 100,000€
Are you the owner of your primary residence, or do you live with parents who own theirs?	Yes No
What is the amount of your rent? (0 if you own or live with parents who own.)	<i>Numeric</i>
What is the estimated value of your real estate assets (or your family's if you are a student)?	<i>Numeric</i>

What is the estimated value of your financial assets (or your family's if you are a student)?	<i>Numeric</i>
How much are you able to save each month?	<i>Numeric</i>
Do you currently have any loans?	Yes No
What is the estimated total of your loans? (0 if none.)	<i>Numeric</i>
Financial Knowledge & Experience	
If you wanted to invest now, what would be the primary goal of your investment?	Grow my savings Prepare a major purchase Save for emergencies Plan for retirement
If you wanted to invest now, what would be your investment horizon?	Less than 5 years 5–10 years 10–15 years More than 15years
“A high expected return implies a high risk of capital loss.” Do you believe this statement is true?	True False I don't know
What is the name of the payment a company makes to its shareholders?	Dividend Interest Profit I don't know
“The more volatile a financial asset is. . .”	The riskier it is The less risky it is I don't know
What is a share (stock)?	A property title

	A debt title
	An investment fund
	I don't know
Have you ever experienced losses on your financial investments?	No
	Yes, up to 10%
	Yes, up to 20%
	Yes, more than 20%
If your investment loses 10% of its value in 3 months, what would you do?	Reinvest to seize the opportunity
	Hold without panicking
	Sell some to limit losses
	Sell all
	I don't know
Have you ever placed money in a life insurance contract, a brokerage account, or an equity savings plan (PEA)?	Yes
	No

Table 1.A.3: Digital Knowledge and Privacy Measures

Question	Possible answer(s)
Digital Knowledge (Europass)	
I know how to verify that a website asking for personal data is secure (e.g., https, security badge).	I don't know how to do this
	I can do it with help
	I can do it by myself
	I can do it confidently and, if necessary, guide others.
When faced with a technical problem, I can find solutions on the Internet.	I don't know how to do this
	I can do it with help

	I can do it by myself I can do it confidently and, if necessary, guide others.
I know how to use online learning tools to improve my digital skills (e.g., video tutorials, online courses).	I don't know how to do this I can do it with help I can do it by myself I can do it confidently and, if necessary, guide others.
I know how to protect myself from unwanted or malicious messages online (e.g., spam, phishing emails).	I don't know how to do this I can do it with help I can do it by myself I can do it confidently and, if necessary, guide others.
I know how to copy and move files (documents, images, videos, etc.) between folders, devices, or cloud storage.	I don't know how to do this I can do it with help I can do it by myself I can do it confidently and, if necessary, guide others.
I know how to create a profile in various digital environments for personal or professional purposes.	I don't know how to do this I can do it with help I can do it by myself I can do it confidently and, if necessary, guide others.
I know how to create and edit digital text files (e.g., Word, OpenDocument, Google Docs).	I don't know how to do this

	I can do it with help I can do it by myself I can do it confidently and, if necessary, guide others.
When using a search engine, I can take advantage of its advanced features.	I don't know how to do this I can do it with help I can do it by myself I can do it confidently and, if necessary, guide others.
I know how to use advanced features in a video-conference (e.g., moderation, recording).	I don't know how to do this I can do it with help I can do it by myself I can do it confidently and, if necessary, guide others.
I know how to create new things by combining different types of content (e.g., text and images).	I don't know how to do this I can do it with help I can do it by myself I can do it confidently and, if necessary, guide others.
Privacy Psychometrics Measures (Dienlin & Trepte, 2015)	
Informational Privacy	
How much personally identifying information have you already posted on your social networks?	None Very little A little Moderate Average High Very high

How much personally identifying information can generally be found about you on your social networks?

None

Very little
A little
Moderate
Average
High
Very high

How accurately can strangers identify you on your social networks?

None

Very little
A little
Moderate
Average
High
A lot

How much personally identifying information do you currently want to share on your social networks?

None

Very low
Low
Moderate
Average
High
Very high

How much personally identifying information would you generally like to see on your social networks?

None

Very low
Low
Moderate
Average

	High
	Very high
How accurately would you like to be identifiable by strangers on your social networks?	None
	Very low
	Low
	Moderate
	Average
	High
	Very high
I think sharing identifying information on my social networks is:	Not useful
	Slightly useful
	Rather not useful
	Neutral
	Rather useful
	Useful
	Very useful
I think sharing identifying information on my social networks is:	Very disadvantageous
	Disadvantageous
	Rather disadvantageous
	Neutral
	Rather advantageous
	Advantageous
	Very advantageous
I think sharing identifying information on my social networks is:	Very worrying
	Worrying
	Rather worrying
	Neutral
	Rather reassuring

	Reassuring
	Very reassuring
I think sharing identifying information on my social networks is:	Very dangerous
	Dangerous
	Rather dangerous
	Neutral
	Rather not dangerous
	Slightly dangerous
	Not dangerous
I think sharing identifying information on my social networks is:	Very imprudent
	Imprudent
	Rather imprudent
	Neutral
	Rather not imprudent
	Slightly imprudent
	Not imprudent
I think sharing identifying information on my social networks is:	Very bad
	Bad
	Rather bad
	Neutral
	Rather good
	Good
	Very good
Social Privacy	
To what extent do you currently limit access to your social network profiles?	Not at all
	Very little
	A little
	Moderately

	Average
	A lot
	Extremely
To what extent is the visibility of your social network content restricted?	Not at all
	Very little
	A little
	Moderately
	Average
	A lot
	Extremely
To what extent are your social network profiles restricted for specific people?	Not at all
	Very little
	A little
	Moderately
	Average
	A lot
	Extremely
To what extent would you like to restrict access to your social network profiles currently?	Not at all
	Very little
	A little
	Moderately
	Average
	A lot
	Extremely
To what extent would you like the visibility of your social network content to be limited?	Not at all
	Very little
	A little
	Moderately

	Average
	A lot
	Extremely
To what extent would you like access to your social network profiles restricted for certain people?	Not at all
	Very little
	A little
	Moderately
	Average
	A lot
	Extremely
I think restricting access to one's social profile is:	Not useful
	Slightly useful
	Rather not useful
	Neutral
	Rather useful
	Useful
	Very useful
I think restricting access to one's social profile is:	Very disadvantageous
	Disadvantageous
	Rather disadvantageous
	Neutral
	Rather advantageous
	Advantageous
	Very advantageous
I think restricting access to one's social profile is:	Very worrying
	Worrying
	Rather worrying
	Neutral
	Rather reassuring
	Reassuring
	Very reassuring

I think restricting access to one's social profile is:

Very unpleasant

Unpleasant

Rather unpleasant

Neutral

Rather not unpleasant

Slightly unpleasant

Not unpleasant

I think restricting access to one's social profile is:

Very mean

Mean

Rather mean

Neutral

Rather fair

Fair

Very fair

I think restricting access to one's social profile is:

Very bad

Bad

Rather bad

Neutral

Rather good

Good

Very good

Psychological Privacy

To what extent do you share personal things on your social networks?

Very impersonal

Rather impersonal

Slightly impersonal

Neutral

Slightly personal

Rather personal

Very personal

How accurately do your social network profiles reflect your personality?

Very impersonal

	Rather impersonal
	Slightly impersonal
	Neutral
	Slightly personal
	Rather personal
	Very personal
To what extent are your social network profiles personal?	Very impersonal
	Rather impersonal
	Slightly impersonal
	Neutral
	Slightly personal
	Rather personal
	Very personal
How much personal information do you wish to communicate on your social networks?	None
	Very low
	Low
	Moderate
	Average
	High
	Very high
To what extent do you want your social network profiles to resemble your full personality?	Not at all
	Very little
	A little
	Moderately
	Average
	A lot
	Extremely
To what extent do you want your social network profiles to be personal?	Not at all

	Very little
	A little
	Moderately
	Average
	A lot
	Extremely
I think sharing personal information on social networks is:	Not useful
	Slightly useful
	Rather not useful
	Neutral
	Rather useful
	Useful
	Very useful
I think sharing personal information on social networks is:	Very disadvantageous
	Disadvantageous
	Rather disadvantageous
	Neutral
	Rather advantageous
	Advantageous
	Very advantageous
I think sharing personal information on social networks is:	Very worrying
	Worrying
	Rather worrying
	Neutral
	Rather reassuring
	Reassuring
	Very reassuring
I think sharing personal information on social networks is:	Very dangerous

	Dangerous
	Rather dangerous
	Neutral
	Rather not dangerous
	Slightly dangerous
	Not dangerous
I think sharing personal information on social networks is:	Very unpleasant
	Unpleasant
	Rather unpleasant
	Neutral
	Rather not unpleasant
	Slightly unpleasant
	Not unpleasant
I think sharing personal information on social networks is:	Very bad
	Bad
	Rather bad
	Neutral
	Rather good
	Good
	Very good
Privacy Concerns	
In general, how concerned are you about your privacy when using the Internet?	Very strongly
	Strongly
	Moderately
	A little
	Not at all
Are you concerned that online organizations are not who they claim to be?	Very strongly
	Strongly

	Moderately
	A little
	Not at all
Are you concerned that you are asked for too much personal information when registering or shopping online?	Very strongly
	Strongly
	Moderately
	A little
	Not at all
Are you concerned about online identity theft?	Very strongly
	Strongly
	Moderately
	A little
	Not at all
Are you concerned that people online are not who they claim to be?	Very strongly
	Strongly
	Moderately
	A little
	Not at all
Are you concerned that information about you could be found on an old computer?	Very strongly
	Strongly
	Moderately
	A little
	Not at all
Are you concerned that strangers might obtain personal information about you from your online activities?	Very strongly
	Strongly
	Moderately

	A little
	Not at all
Are you concerned that a message you send online might be read by someone other than its intended recipient?	Very strongly
	Strongly
	Moderately
	A little
	Not at all
Are you concerned that messages you receive online might not come from the person they claim to be?	Very strongly
	Strongly
	Moderately
	A little
	Not at all
Specific Privacy Behaviors	
On my social networks, I use my real first name (exactly as on my passport).	No
	Yes
On my social networks, I use my real last name (exactly as on my passport).	No
	Yes
On my social networks, I display my current address.	No
	Yes
On my social networks, I display my phone number even when it is not required.	No
	Yes
Have you ever posted a religious, political, or ethical statement on social networks?	No

	Yes
How often do you post on social networks?	Never Once every few months Once a month Several times a month Several times a week Once a day Several times a day
Feedback Items	
During this experiment, how accurate were the questionnaire responses you provided?	Not at all A little Moderately Precise Very precise
Have you ever experienced a personal data leak?	Yes, once Yes, several times No
During this experiment, were you interested in other participants' data leaks?	Not at all A little Moderately Strongly Very strongly
How would you characterize the risk of a data leak in this experiment?	Very dangerous Dangerous Rather dangerous Neutral Rather not dangerous Slightly dangerous Not dangerous

Did the risk of a data leak influence your investment decisions?	Not at all
	A little
	Moderately
	Strongly
	Very strongly

Are you more aware of potential risks associated with data leaks after participating in this experiment?	Not at all
	Slightly
	Moderately
	Strongly
	Extremely

Instructions

Table 1.A.4: Instructions

Thank you for your participation!

You are about to participate in a decision-making study. Each participant receives exactly the same instructions. All decisions and interactions are anonymous.

It is important that you remain silent and do not communicate in any way with other participants outside of your interactions in the experimental interface.

If you have any questions, please raise your hand and I will come to answer them personally.

You will be paid in cash at the end of the experiment; your earnings will depend on your decisions and chance.

Remuneration:

- All participants earn a fixed payment of €5 for their participation.
- You will receive additional compensation for three experimental tasks: two initial tasks and one main experimental task. Payments for these tasks will be communicated at the end of the experimental session.
- Total compensation will be the sum of your fixed payment and your payments for the three tasks.
- The two initial tasks will be performed and remunerated directly in euros.
- The main experimental task will be conducted using experimental monetary units called ECUs. ECUs earned will be converted at a rate of 12.5 ECUs per €1.

Instructions for the main task:

Allocation In each decision interface, you will have an allocation of 100ECUs.

Investment choice You must choose to invest in either Option A or Option B. Both options involve risky investment opportunities.

Investment amount You will be asked to enter the amount you wish to invest in the chosen option. Any units not invested will be added to your potential gains or losses.

Investment scenarios Each option is characterized by three scenarios, with probabilities displayed in the interface:

- 25% chance of a favorable scenario,
- 50% chance of a realistic scenario,
- 25% chance of an unfavorable scenario.

The scenario implemented will determine the corresponding payoff. Probabilities remain constant, but payoffs may change across interfaces.

Expected payoff The interface shows the expected payoff (the average return) for each option.

Payoff difference The interface also shows the difference in expected payoff between the two options, in percentage points.

Risk Both Option A and Option B have the same volatility (financial risk).

Probability of data leak One option is randomly assigned a probability of personal data leak:

- If Option A has a leak probability, Option B does not, and vice versa—this assignment is common to all participants.
- Leak probabilities are independent of expected payoffs and vary randomly from 0% to 40% in 5% increments.

Data leak outcome A leak occurs if the random draw falls below the leak probability. The result page will display leaked personal data (age and pseudonym plus two randomly chosen economic variables) for all participants whose chosen option experienced a leak.

Result page Your decision outcomes will appear on a result page once all participants have decided.

Number of decisions You will face 60 decision interfaces. One of your 60 decisions will be randomly selected for payment.

1.6.2 Additional Results

Table 1.A.5: Summary Statistics of Continuous Variables

	Mean	SD	Min	Median	Max
Age	28.2	7.6	19.0	26.0	52.0
<i>Economic Variables</i>					
Rent	466.9	479.1	0.0	453.5	3750.0
Property Asset	126355.6	226042.6	0.0	462.5	1300000.0
Financial Wealth	39779.1	121760.3	0.0	7500.0	1000000.0
Savings	217.5	276.8	0.0	100.0	1200.0
Indebtedness	7386.2	28955.8	0.0	0.0	200000.0
<i>Consistency Ratios</i>					
Not Included in Leaks	96.4	8.5	42.9	100.0	100.0
Included in Leaks	82.5	15.1	25.0	87.5	100.0

Table 1.A.6: Summary Statistics of Categorical Variables

		N	%
<i>Gender</i>	Women	91	48.9
	Men	95	51.1
<i>Marital</i>	Single	148	79.6
	In a relationship	22	11.8
	Divorced	2	1.1
	Married	11	5.9
	Prefer not to answer	3	1.6
<i>Children</i>	1 child	7	3.8
	2 children	5	2.7
	3 children or more	3	1.6
	None	170	91.4
	Prefer not to answer	1	0.5
<i>Socio-professional status</i>	Craftsperson, shopkeeper, and business owner	7	3.8
	Other	10	5.4
	Executive and higher intellectual profession	20	10.8
	Employee	31	16.7
	Student	106	57.0
	Intermediate profession	12	6.5
<i>Education</i>	PhD	3	1.6
	Bachelor's degree or Advanced Technician's Certificate	68	36.6
	Master's degree	83	44.6
	Primary or no education	2	1.1
	Secondary education	30	16.1
<i>Annual Salary</i>	0-10 000 euros	105	56.5
	10 000-25 000 euros	30	16.1
	25 000-45 000 euros	42	22.6
	45 000-60 000 euros	6	3.2
	60 000-100 000 euros	2	1.1
	More than 100 000 euros	1	0.5

Table 1.A.7: Summary Statistics of Data included in Data Leaks

Measure	Pre-registration Questionnaire	Experimental Questionnaire	95% CI	<i>p</i> -value
Annual Salary (Cat.)	1.81	1.79	[−0.190; 0.243]	0.808
Homeownership (Yes or No)	0.278	0.288	[−0.102; 0.081]	0.819
Rent (Cont.)	471.73	466.81	[−92.31; 102.15]	0.921
Property Asset (Cont.)	121693.9	125679.9	[−50601.70; 42629.56]	0.867
Financial Asset (Cont.)	32797.94	39673.32	[−29675.04; 15924.28]	0.553
Savings (Cont.)	218.30	218.98	[−59.46; 58.10]	0.982
Indebted (Yes or No)	0.171	0.155	[−0.059; 0.091]	0.676
Indebtedness (Cont.)	10749.87	7346.68	[−5763.51; 12569.88]	0.465

Table 1.A.8: Consistency of responses between pre-registration and experimental questionnaire

Measure	Mean	Standard Deviation
Overall	82.55%	0.151
Annual Salary (Cat.)	89.84%	0.303
Homeownership (Yes or No)	93.58%	0.246
Rent (Cont.)	80.21%	0.399
Property Asset (Cont.)	72.19%	0.449
Financial Asset (Cont.)	63.10%	0.484
Indebted (Yes or No)	97.33%	0.162
Indebtedness (Cont.)	90.37%	0.296

Table 1.A.9: Determinants of Allocation Choices: Economic and Behavioral Characteristics

	<i>Dependent variable: Data Leak Allocation</i>					
	(1)	(2)	(3)	(4)	(5)	(6)
Digital Knowledge	-0.046 (0.036)	-0.043 (0.036)	-0.053 (0.036)	-0.049 (0.036)	-0.034 (0.038)	-0.031 (0.038)
<i>Behavioral characteristics</i>						
Loss Aversion		-0.129 (0.085)		-0.106 (0.112)		-0.150 (0.112)
Risk Aversion			-0.093 (0.087)	-0.041 (0.108)		-0.022 (0.108)
<i>Economic characteristics (pre-registered)</i>						
Income: 0–10k					Ref.	Ref.
10–25k euros					-0.143 (0.298)	0.014 (0.298)
25–45k euros					-0.084 (0.412)	0.017 (0.412)
45–60k euros					-0.360 (0.650)	-0.474 (0.650)
60–100k euros					-0.058 (0.603)	0.349 (0.603)
>100k euros					-0.973 (0.817)	-1.401 (0.817)
Homeowner					-0.351 (0.280)	-0.314 (0.280)
Rent					0.063 (0.113)	0.044 (0.113)
Property assets					0.263* (0.129)	0.274* (0.129)
Financial wealth					-0.035 (0.057)	-0.059 (0.057)
Savings					0.051 (0.136)	0.089 (0.136)
Indebtedness					0.149 (0.174)	0.071 (0.174)
Constant	2.928 (1.951)	3.473 (2.017)	3.582 (2.280)	3.543 (2.236)	3.470 (1.959)	3.991* (1.959)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Num. obs.	5231	5231	5231	5231	5231	5231
Log likelihood	-2080.06	-2070.16	-2068.73	-2062.97	-2059.81	-2043.02
Pseudo R ² (McFadden)	0.239	0.243	0.244	0.246	0.247	0.253

Notes. Logistic regressions of the choice to allocate to the data-leak option. The sample is restricted to rounds with strictly positive data-leak probability and non-negative spreads. Standard errors (in parentheses) are clustered at the participant level. Experimental parameters (probability and spread) are included but not reported. Controls include age, gender, number of children, marital status, education, professional category, financial literacy (PCA index), and response precision.

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Chapter 2

Chapter 2

Algorithmic Transparency and Portfolio Choices: Field Evidence

Co-written with Béatrice Boulu-Reshef & Alexis Direr

2.1 Introduction

Investment in financial assets, such as equities, fixed-income securities, or other instruments, constitutes a fundamental mechanism for wealth accumulation. Households may engage in these investments either directly or indirectly through vehicles like retirement plans. However, the allocation of financial resources under uncertainty remains one of the most complex decisions in household investment management. A well-documented phenomenon in this context is the “stock market participation puzzle” (Guiso, Jappelli, & Terlizzese, 1996), which refers to the persistent reluctance of a substantial share of households in advanced economies to hold equities, ultimately hindering their long-term wealth accumulation.

Traditionally, households have relied on professional financial advisors to navigate portfolio allocation challenges. However, the evolving structure of financial markets has influenced individuals’ saving and investment behaviors. The increasing complexity of financial products and the proliferation of financial innovations have expanded the range of choices available to investors, sometimes leading to suboptimal decision-making. Over time, interactions between financial advisors and individual investors have evolved, with financial service providers offering an expanding array of advisory services. Consequently, individuals have assumed greater responsibility for their financial well-being (Hung & Yoong, 2013). Behavioral finance literature suggests that, in the absence of professional guidance, investors frequently fail to construct optimal portfolios. Instead, they tend to rely on heuristics and simplistic decision rules when

making investment choices (Agnew & Szykman, 2005; Benartzi & Thaler, 2001). Given that household investment decisions often diverge from the prescriptions of optimal portfolio theory, reliance on heuristics and investment mistakes can lead to significant welfare losses (Dominitz & Hung, 2006).

The emergence of robo-advising — algorithmic, largely automated investment advice and portfolio management delivered via digital platforms — in financial technology (FinTech) has been driven not only by households' lack of specialized financial knowledge but also by the need to mitigate biases inherent in traditional human advising (Foerster, Linnainmaa, Melzer, & Previtero, 2018). A key disruptive feature of robo-advising in asset management is its ability to provide households with investment expertise and simplified decision-making frameworks that would otherwise be inaccessible to median investors with limited financial literacy. Additionally, robo-advisors offer these services at a significantly lower cost compared to traditional human financial advisors, eliminating high advisory fees (D'Acunto & Rossi, 2023). A major advantage of FinTech innovations, including robo-advising, is the reduction in the cost of financial intermediation, driven by technological advancements and increased competition. Philippon (2019) documents that, for the first time in a century, the unit cost of financial intermediation has declined over the past decade. A direct consequence of this cost reduction is enhanced financial inclusion. The expansion of FinTech services has been associated with a decrease in barriers to accessing financial products, particularly for marginalized groups. For mortgage lending, (Bartlett, Morse, Stanton, & Wallace, 2018) show that African-American and Latino borrowers pay higher interest rates than observably similar borrowers, but FinTech lenders exhibit about 40% smaller racial pricing gaps than traditional face-to-face lenders. The integration of robo-advisors—automated online applications that provide financial guidance—has also transformed the wealth management industry. The entry of FinTech firms into this market has altered cost structures by reducing fixed expenses related to salaries and office space, which, in turn, has lowered minimum investment requirements and advisory fees (Abraham, Schmukler, & Tessada, 2019). As a result, Philippon (2019) suggests that this shift in cost dynamics has the potential to increase stock market participation, particularly among relatively less wealthy households.

Despite these advantages, the adoption of robo-advisors remains relatively limited, as they still manage only a small fraction of private assets (Cardillo & Chiappini, 2024). While robo-advisors can lower investment thresholds, reduce advisory costs, and mitigate moral hazard issues inherent in traditional advisor-advisee relationships (Abraham et al., 2019), many investors remain hesitant to entrust their financial decisions to automated systems. Empirical evidence suggests that individuals continue to value in-person financial advisory services provided by banks, preferring direct human interaction over digital solutions (Jung, Dorner, Weinhardt, & Puzmaz, 2018). Automated algorithmic systems and artificial intelligence (AI) are increasingly assisting economic agents in decision-making processes across various domains, including medical diagnostics (Liu, Zhou, & Wang, 2019), financial services (Binns et al.,

2018), fraud detection (Cirqueira, Nedbal, Helfert, & Bezbradica, 2020), and online advertising (Eslami, Krishna Kumaran, Sandvig, & Karahalios, 2018). While these AI-driven systems enhance efficiency and reduce costs, they also introduce challenges related to transparency and interoperability. The field of Explainable AI (XAI) seeks to address these challenges by making AI system outputs more understandable for human users (Adadi & Berrada, 2018). However, much of the existing XAI literature focuses on technical solutions (Anjomshoe, Najjar, Calvaresi, & Främling, 2019; Arrieta et al., 2020), with relatively less emphasis on what constitutes a “good” explanation from a human perspective (Riedl, 2019; Weitz, Schiller, Schlagowski, Huber, & André, 2021). This issue is particularly relevant in financial decision-making, where trust and comprehension are crucial for adoption.

Traditional financial advisors not only provide portfolio recommendations but also explain the reasoning behind their choices. However, financial advisors may be influenced by their own preferences and personal biases (Foerster et al., 2018; Linnainmaa, Melzer, Previtero, & Foerster, 2020; d’Astous, Gemmo, & Michaud, 2022), which can lead to conflicts of interest. Moreover, personalized advisory services are often reserved for high-net-worth clients, leaving a gap in financial guidance for median investors. Robo-advisors attempt to fill this gap by offering a disintermediated, automated, and cost-effective approach to financial advice (Abraham et al., 2019). These systems typically employ self-administered questionnaires to assess an investor’s objectives, risk tolerance, and preferences. The collected responses are processed by an algorithm, which assigns a risk score and determines the portfolio’s risk exposure (Abraham et al., 2019). However, the automation of financial advice comes with trade-offs. One significant drawback is the loss of human interaction, which can undermine trust in the system (Morana, Gnewuch, Jung, & Granig, 2020). Additionally, the lack of transparency in how robo-advisors generate recommendations—often perceived as a “black-box” process—further exacerbates this trust deficit. While robo-advisors provide tailored portfolio allocations, they rarely offer users insight into the reasoning behind their recommendations. Therefore, integrating explainability features within robo-advisors could help mitigate investor reluctance to follow algorithmic advice, thereby increasing adoption rates. If individuals can understand why a given portfolio allocation aligns with their financial situation and preferences, they may develop greater trust in algorithmic advice and be more willing to utilize these services for long-term wealth accumulation. We therefore ask whether profile-based explanations, concise disclosures of the drivers of a recommended risk score benchmarked to similar users, increase acceptance of algorithmic risk profile recommendations on a market-deployed robo-advisor.

We study this question with a randomized controlled trial embedded in the interface of a leading French robo-advisor. The experiment randomized 4,645 saving-plan contracts to one of two conditions: (i) a treatment in which users received graphical, profile-based explanations detailing the relative importance of variables (economic situation, liquidity needs, investment horizon, demographics, risk tolerance) and

benchmarking them to a “typical similar user”; or (ii) a control without any additional information beyond the recommended risk profile. Our primary outcome is acceptance, defined as choosing the recommended risk profile at the decision point. We also track the exact deviation from the recommended profile and engagement metrics. Sessions are observed on desktop and mobile, allowing us to study how the usage context moderates responses.

Three results stand out. First, explanations do not raise acceptance: treatment and control accept at statistically indistinguishable rates. Second, device context matters: desktop sessions exhibit lower baseline adherence than mobile. Third, among desktop sessions that choose a safer-than-recommended portfolio, explanations are associated with larger downward moves; however, this pattern loses significance once platform experience is accounted for. Beyond this subgroup, we find little systematic heterogeneity and no improvement in engagement with the platform. Taken together, profile-based transparency is not a universal lever for adherence, and its effects are context-dependent.

Our paper makes three contributions. First, we provide in-market experimental evidence on a widely advocated transparency intervention, documenting a precise null on acceptance and on deviations. Second, we show that behavior is shaped by both interface and experience: device context (desktop vs. mobile) and platform tenure (first-time vs. returning users) are first-order correlates of adherence—mobile and first-time users accept more, while desktop and returning users explore and deviate more; moreover, the desktop-specific amplification of downward moves is sensitive to controlling for experience. Third, we offer managerial guidance: for deployments seeking higher acceptance, adding profile-based explanations does not guarantee improved adherence; rather, firms should test context-contingent designs (e.g., lighter explanations on desktop at the acceptance step, progressive disclosure, alternative benchmarks) instead of expecting uniform gains from transparency.

We implement a single-feature intervention by mirroring how a human advisor would justify a recommendation, grouping drivers into meaningful categories, and situating the user in a peer reference class, while preserving the product’s decision flow. The randomization was implemented within the production system and analyzed independently by the authors.

Our findings point to a clear null: in this randomized, in-market setting, profile-based explanations do not measurably change acceptance, the direction or magnitude of deviations, or engagement. Apparent interactions in simpler models are not robust once we account for user experience. The systematic patterns we observe, higher adherence on mobile and among first-time users, are driven by device and experience with the platform, not by the explanation itself.

The remainder of the paper proceeds as follows. Section 2.2 situates our study within research on explainability, robo-advising, and digital finance. Section 2.3 details the experimental design and measurement. Section 2.4 reports the main effects and robustness. Section 2.5 concludes with implications

for the design of explainable robo-advice.

2.2 Related work

2.2.1 Explainable Artificial Intelligence

The increasing complexity and opacity of many machine learning (ML) systems, particularly those that involve artificial intelligence (AI), have prompted widespread concern about their lack of interpretability (Castelvecchi, 2016). For both end-users and developers, understanding how these systems arrive at their predictions can be challenging. This has led to growing demands for transparency, voiced by stakeholders (Preece et al., 2018), regulators,¹ and users alike. In response, the field of Explainable AI (XAI) has emerged with the goal of making AI-generated outcomes more comprehensible to human users (Adadi & Berrada, 2018).

The objectives of XAI are multifaceted. Beyond promoting algorithmic transparency (Cheng et al., 2019), XAI seeks to clarify the rationale behind individual model predictions for both users and stakeholders (Cai et al., 2019), and ultimately to cultivate trust in AI-driven systems (Shin, 2021; Weitz et al., 2021). A key component of this effort lies in developing frameworks for what constitutes a “good” explanation from a human-centered perspective. For instance, Riedl (2019) emphasizes that an effective XAI system should also account for edge cases—i.e., when outputs appear unintuitive or erroneous. Similarly, Weitz et al. (2021) investigate the potential of virtual agents as mediators of explanation, showing that users often prefer interactive, conversational forms of explanation that mimic human behavior.

The technical implementation of explainability features has been the focus of several contributions. Anjomshoe et al. (2019) outline a three-phase model: generating explanations, communicating them effectively, and evaluating how they are received by users. Meanwhile, Arrieta et al. (2020) distinguish between models that are inherently interpretable and those that require post-hoc interpretability tools. For instance, decision trees are often considered self-explanatory, whereas neural networks typically demand auxiliary methods to make sense of their outputs. In a complementary contribution, Beaudouin et al. (2020) propose a technical framework structured around contextual variables, algorithmic strategies, and explanatory dimensions. Collectively, this body of work defines the core goals and methodologies of the XAI field. For clarity, we refer to this strand of research as “explainability” in the remainder of the paper.

A widely used approach within this literature is feature importance attribution, which aims to highlight the most influential inputs driving a particular decision. These explanations often follow a general format such as: “This instance was classified as category Y due to strong values in features A and B, despite the

¹Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (GDPR).

presence of feature C.” This logic is commonly applied in recommender systems as well, where explanations typically read: “Option X is recommended because of attributes A and B, although attribute C may be less favorable” (Nunes & Jannach, 2017).

Another central distinction in the XAI literature concerns local vs. global explanations (Speith, 2022). Local methods seek to clarify the classification of a single instance, providing insight into why the model produced a specific output for a given input. A prominent example is the LIME algorithm developed by Ribeiro, Singh, and Guestrin (2016a), which build simplified surrogate models to approximate the decision boundary in the local neighborhood of a prediction. In contrast, global explanation techniques aim to describe the overall behavior of the model by identifying general patterns in feature influence across many instances (Speith, 2022; Krzyżiński, Spytek, Baniecki, & Biecek, 2023).

Among global methods, perturbation-based techniques such as SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017; S. B. Cohen, Dror, & Ruppin, 2005) and SAGE (Shapley Additive Global Importance) (Covert, Lundberg, & Lee, 2021) have gained prominence. These methods quantify feature relevance by systematically modifying or removing input variables and observing their effect on model predictions. They generate explanations of the form: “The model assigns items to class Y largely due to features A and B.” SHAP, in particular, is versatile, capable of offering both local and global insights depending on how it is applied. This dual capability makes it especially valuable in sensitive domains such as credit scoring, where understanding how changes in financial attributes affect prediction outcomes is essential. We next turn from methods for generating explanations to evidence on how such explanations shape user understanding, acceptance, and trust.

2.2.2 Explanations

This subsection synthesizes experimental and HCI evidence on the behavioral consequences of providing explanations. Following the distinction proposed by Weitz et al. (2021), we separate two strands within the broader XAI literature. The first, which we refer to as the “explainability” literature, includes the conceptual and technical frameworks detailed earlier in Section 2.2.1, focusing on how to construct and implement interpretable models. The second, which we label the “explanations” literature, investigates the consequences of providing explanations to users and explores their effects across a variety of applications. Since its origins, XAI has expanded to influence numerous domains that rely on artificial intelligence and machine learning (Beaudouin et al., 2020), with particularly active contributions from the fields of Computer Science (CS) and Human-Computer Interaction (HCI).

A prominent line of research in this second strand examines how explanations shape human interaction with algorithms. For example, Cai et al. (2019) explore the role of example-based explanations in image recognition systems. Their findings suggest that such explanations generally enhance user understanding,

though the efficacy depends on the specific type of example presented. Similarly, Shin (2021) evaluates how different explanation models affect the comprehension of non-expert users. While explanations improve both objective and self-reported understanding, the study finds no significant difference between white-box and black-box explanatory styles, and notably, user confidence in the algorithm does not increase alongside understanding. By contrast, Weitz et al. (2021) show that integrating virtual agents capable of explaining an audio classification system can increase user confidence, although the size of the effect is contingent on the agent’s complexity and feature set.

Other studies apply explanation strategies in domains such as social media and digital advertising. Rader et al. (2018) investigate user reactions to various ways of explaining Facebook’s News Feed algorithm, finding that all forms of explanation improve users’ understanding and help them recognize algorithmic bias. Likewise, Eslami et al. (2018) show that providing explanations for targeted advertising algorithms improves user comprehension, while also fostering “algorithmic disillusionment”—a shift in beliefs about how well the algorithm performs. In the context of recommender systems, experimental studies by Herlocker et al. (2000) and Cramer et al. (2008) demonstrate that explanations for recommendations increase user acceptance, especially when users are informed about the reasons underlying a particular suggestion (e.g., similarities with other users’ behavior).

Taken together, this body of work suggests that explanatory interventions can play a meaningful role in demystifying algorithmic processes for non-technical users, thereby increasing understanding, acceptance, and in some cases, trust. However, the magnitude and direction of these effects are highly context-dependent, varying according to the explanation strategy, the system’s complexity, and the domain of application.

Importantly, the observed relationship between explanation and trust has implications beyond HCI and CS. It intersects with insights from the economic literature on decision delegation. This literature emphasizes the role of trust in shaping individuals’ willingness to delegate choices to third parties. Theoretical contributions suggest that confidence in a decision-making agent increases the likelihood of delegation (Bloom et al., 2012; Gur & Bjørnskov, 2017), a finding supported by empirical evidence across various sectors, including finance, labor, and organizational management (Guiso et al., 2008; Cingano & Pinotti, 2012; Özer et al., 2018; Holzmeister et al., 2023). Thus, XAI’s potential to enhance trust via transparent explanations may have important consequences for how individuals interact with algorithmic decision-makers in economic contexts. Taken together, these studies show that the impact of explanations is highly context-dependent. We therefore narrow our focus to robo-advising, where delegation and trust are first-order and regulatory transparency is salient (Section 2.2.3).

2.2.3 XAI, explanations, and robo-advisors

We situate explanations within robo-advising, linking trust, delegation, and regulatory requirements to concrete explanation designs.

In the context of robo-advising, the integration of Explainable AI (XAI) has emerged as a critical factor in enhancing user trust, regulatory compliance, and overall adoption. While AI and ML offer powerful tools for automating portfolio management and optimizing financial recommendations, their inherent opacity can create significant barriers to user acceptance (Patel & Lincoln, 2019; Mestikou, Smeti, & Hachaïchi, 2023). As noted by Bianchi and Brière (2020), trust plays a fundamental role in the delegation of financial decisions, and studies have shown that algorithm aversion (Dietvorst et al., 2015; Dietvorst, Simmons, & Massey, 2018) —where users prefer human judgment over AI, even when AI performs better—remains a challenge in automated financial services. However, recent experimental research suggests that providing adequate explanations can mitigate these concerns and, in some cases, even increase the adoption of algorithmic advice over human recommendations.

Ben David, Resheff, and Tron (2021) conducted an experimental study using Amazon Mechanical Turk to examine the effects of different types of explanations on users' perception and adoption of financial advisory algorithms. In a simulated financial decision-making scenario, participants were exposed to advisory systems with varying levels of explanation: “no explanation,” “human expert,” “global,” “feature-based,” and “performance-based.” Their findings challenge the conventional view of algorithm aversion, showing that users were more inclined to adopt AI-generated advice over human recommendations when explanations were provided. In particular, accuracy-based explanations in the initial phases led to higher adoption rates, while feature-based or performance-based explanations were especially effective in mitigating trust erosion in cases of algorithmic failure.

Similar insights were found in a study by Deo and Sontakke (2021), who examined how users interact with explanations in a robo-advisory setting. The authors developed a replica robo-advisor system that categorized users into risk profiles based on their responses to ten questions, subsequently recommending funds matched to their risk category through a recommender system. Users received explanations through a SHAP-like visualization, illustrating both local and global feature importance in determining their financial recommendations. Their findings emphasize that explanations improve user comprehension and engagement, particularly when they explicitly link user inputs to system outcomes. However, they also highlight the importance of simplicity and clarity, as overly complex explanations risk disrupting the user experience rather than enhancing trust and adoption.

These findings reinforce the broader literature on the role of XAI in financial decision-making. While personalization is a key advantage of robo-advisors, excessive reliance on AI-driven individualization can

introduce risks related to model uncertainty and measurement errors in user preferences (Capponi, Olafsson, & Zariphopoulou, 2022). As Bianchi and Briere (2021) emphasize, the future of robo-advising may not necessarily lie in more complex AI but in a balanced approach where explainability, transparency, and human-AI interaction are prioritized. Explanations serve not only to increase user trust but also to ensure compliance with regulatory frameworks that emphasize accountability in automated financial services (Fein, 2017; Strzelczyk, 2017; European Commission, 2019). Ultimately, integrating XAI methodologies into robo-advisors has the potential to bridge the gap between algorithmic decision-making and user expectations, fostering broader adoption and trust in financial AI systems. Yet the effectiveness of explanations also depends on how users access advice. Device and interface choices may shift cognition and behavior, so we place robo-advising within the broader digital-finance environment (Section 2.2.4).

2.2.4 Digital finance

We use this literature to motivate the role of the interaction channel in financial decisions—an element that will guide our experimental design and heterogeneity analysis by device. Digital finance platforms, encompassing mobile apps, online portals, robo-advisors, and AI-driven tools, fundamentally change not only the availability of financial products but also the process by which individuals make decisions. By embedding decision support, minimizing transaction friction, and shifting interactions onto smartphones, digital finance can induce behavioral shifts in decision-making.

Empirical works find that exposure to digital products and services correlates with higher household allocation of financial assets in risky assets. Most studies suggest that digital finance can enhance households’ risk-taking capabilities. Households are exposed to various financial platforms such as mobile payments and online shopping, and may experience less risk aversion and increased risk-bearing levels (Hong, Lu, & Pan, 2020). Digital finance may stimulate more individuals to channel funds into risky financial assets and increase the number of households invested in such assets (Shen, Hu, & Zhang, 2022). Finally, recent empirical evidence from China demonstrates that higher levels of digital financial development are positively associated with household risk-taking (Hu, Guo, Shang, & Zhang, 2024). Using the Digital Financial Inclusion Index (DFII) and microdata from the China Household Finance Survey, the authors find that a one-percent increase in digital finance corresponds to a 0.13 percent rise in the share of risky assets in total household portfolios, and similarly increases the likelihood that a household holds any risky financial assets.

However, a significant portion of the uptick in household risk-taking can be attributed not merely to increased access or literacy but to the cognitive context in which decisions occur. Smartphone environments create a distinct “mobile mindset” (Lurie et al., 2018) that favors fast, intuitive processing over slower, deliberative thinking (Kahneman, 2013). In a controlled lab experiment with random assign-

ment to device, participants using smartphones exhibited a present-bias parameter roughly six percentage points lower than computer users, indicating heightened impulsivity, and were about twelve points less willing to incur time costs to search for payoff-relevant information (Mograbi, 2022). Moreover, habitual smartphone users similarly demonstrate a greater reliance on automatic, heuristic-based processing, opting for default or recommended options without systematically comparing alternatives, compared to when they use larger-screen devices (R. J.-H. Wang, Malthouse, & Krishnamurthi, 2015).

2.3 Experimental design

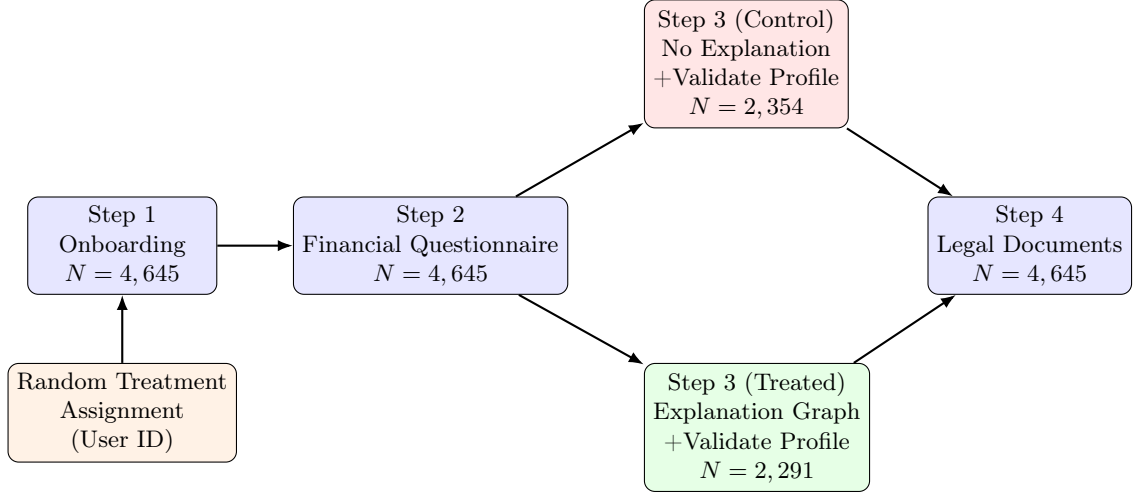
2.3.1 Onboarding flow (“tunnel”)

In our setting, each prospective client who engages with the robo-advisor completes a standardized, multi-step onboarding process. For consistency with the partner’s internal terminology², we refer to this flow as the “tunnel” when discussing specific screens and steps. This structured sequence consists of four distinct stages, referred to as “Steps”, through which clients progress sequentially. Each step serves a specific function in the investment decision-making and account setup process (see Figure 2.1). In Step 1, a unique User ID is assigned and linked to the email account used to register, and the step concludes when the client initiates an investment simulation. In Step 2, the client completes a comprehensive financial questionnaire, which serves as the foundation for the risk assessment and portfolio recommendation process. Submission of this questionnaire simultaneously generates a new Contract ID linked to the corresponding User ID. Step 3 comprises the validation of the proposed risk profile, investment vehicle selection, and portfolio allocation against the client’s stated objectives and constraints. At Step 3, users may open a “Modify” panel listing alternative risk profiles with side-by-side allocations. Acceptance is recorded immediately after closing this panel when the final profile is confirmed and matches the originally recommended profile. Finally, Step 4 requires submission of all requisite legal documentation and the formal execution of the contractual agreement, thereby completing account opening with the robo-advisor.

A key component of this onboarding process is the financial questionnaire, which gathers comprehensive information regarding the client’s financial situation, investment preferences, and constraints. Specifically, the questionnaire includes inquiries on liquidity needs, risk tolerance, financial literacy, investment horizon, and investment objectives. Additionally, it collects various economic indicators such as financial wealth, income levels, savings capacity, and the amount allocated for the initial investment.

²The industry partner’s identity is anonymized under an NDA that was a precondition for data access. Owing to confidentiality constraints and the partner’s deployment timeline, the study was not preregistered. The intervention and its implementation were co-designed with the partner to ensure product feasibility; however, the authors retained full independence over the empirical analysis and the write-up. The partner had no editorial control and no right to approve or veto publication.

Figure 2.1: Flowchart of the Experiment



Furthermore, socio-demographic variables are also recorded to provide a more holistic understanding of the investor profile. Upon completion of the questionnaire, the collected data are processed through a proprietary algorithm designed to generate a risk score. This score is instrumental in determining the client’s risk profile, which subsequently dictates the composition of their recommended investment portfolio. The algorithm not only assesses the appropriate level of risk exposure but also identifies the optimal investment vehicle. For the purposes of our study, we restrict attention to a tax-efficient saving plan (*assurance-vie*), the platform’s flagship product through which most clients invest. Focusing on a single investment vehicle ensures a common decision environment and maximizes sample size.

Within the context of “*assurance-vie*” products, the generated risk score ranges from 2 to 10. A risk score of 2 corresponds to a highly conservative portfolio, allocated mostly to “Fonds Euro” (see Table 2.1). This investment vehicle provides capital protection and is characterized by low risk, albeit with limited return potential. At the opposite end of the risk spectrum, a score of 10 represents a fully equity-based portfolio, composed entirely of stocks, thus entailing a significantly higher level of risk. Intermediate risk profiles, with scores between 2 and 10, are defined by varying allocations between “Fonds Euro”, bonds, and equities. As the risk score increases, the proportion of “Fonds Euro” decreases, while the allocations to bonds and equities rise accordingly, reflecting a progressive shift towards a more aggressive investment strategy. Importantly, after receiving their assigned risk profile, prospective clients have the option to access a “modify option”, which allows them to compare their recommended profile against other available risk profiles. If desired, they may adjust their final risk selection before proceeding with investment validation and contract signing.

The intervention is an A/B test with randomization at Step 1 and treatment display at Step 3, immediately after completion of the Step 2 questionnaire. The treatment group receives an explanation

Table 2.1: Target asset allocation by risk profile (in %)

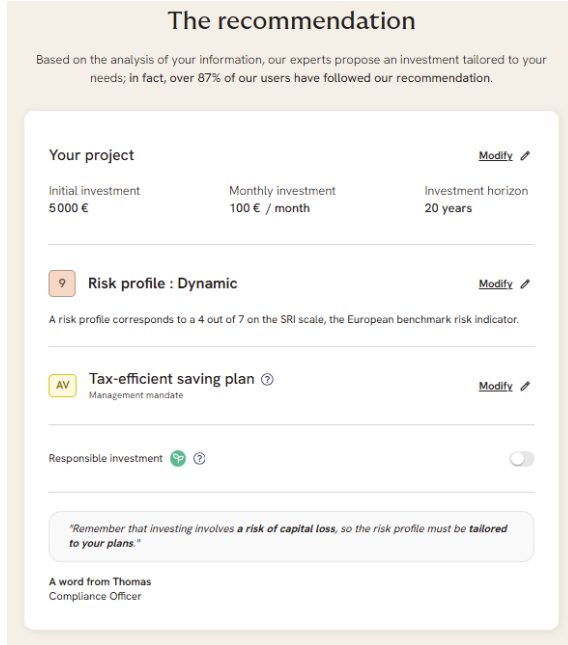
Asset Class	Risk Profiles								
	2	3	4	5	6	7	8	9	10
Euro Funds	70	60	40	20	0	0	0	0	0
Bond ETFs	15	20	30	40	50	40	30	20	0
Stock ETFs	15	20	30	40	50	60	70	80	100

of the assigned risk profile, which focuses solely on the rationale behind the score generated by the algorithm. In contrast, the baseline group does not receive any explanatory information. The treatment is randomly assigned at the beginning of Step 1 based on the User ID. Consequently, if a prospective client undergoes the onboarding process multiple times, a frequent occurrence in practice, they will consistently receive the same treatment condition. This design ensures that exposure to the intervention remains stable, thereby minimizing selection bias and preserving the integrity of the experimental framework. Randomization achieved balance across pre-treatment covariates; Mann–Whitney and χ^2 tests fail to reject equality across arms (all $p > 0.11$; Bonferroni-adjusted $p \geq 0.47$, see Table 2.A.1 in Appendix 2.6.2).

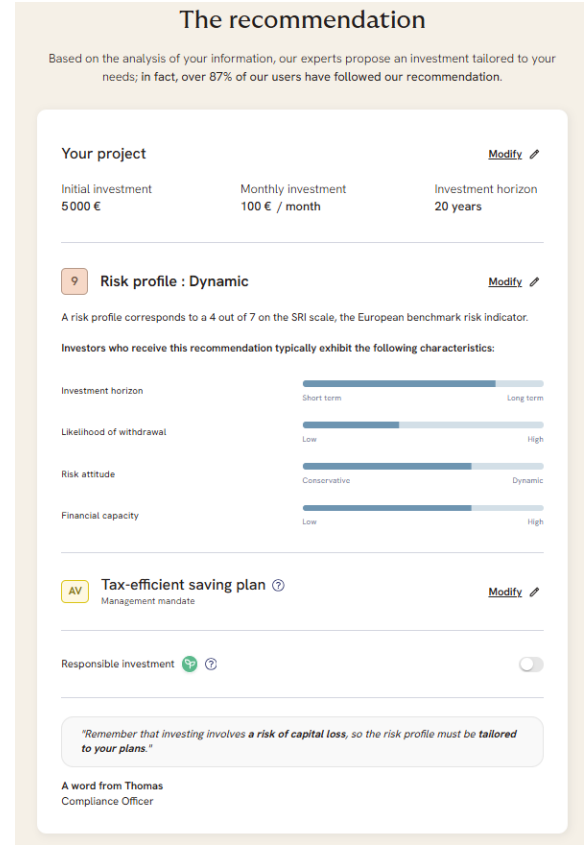
2.3.2 Explanation Methodology

In this study, we focus on the transparency of the algorithm’s recommendations generated by a financial algorithm embedded within a robo-advisor. The algorithm computes a risk profile score for prospective clients, which in turn informs the investment recommendations they receive. To improve transparency and user understanding of the profile assessment process, we implement a graphical representation that visually communicates the relative importance of various factors contributing to the profile score. Specifically, the bar chart displays normalized weights for key determinants (liquidity constraints, risk preferences, economic characteristics, investment horizon) on a 0–10 scale. For visualization, we truncate displayed bars to the range 1–9 to avoid saturation at extremes while not altering relative rankings. To aid interpretation, the axis is annotated with attribute-specific verbal anchors (e.g., “short-term” and “long-term” for horizon; “low” and “high” for liquidity constraints; “conservative” and “dynamic” for risk attitude; “low” and “high” for financial capacity), which describe the underlying dimension rather than the importance units. An example of the visual explanation shown to treated users is provided in Figure 2.2b. The figure illustrates the characteristics of a user assigned risk profile 9. The accompanying bars show how individuals with similar profiles typically score on attributes such as investment horizon, withdrawal likelihood, risk attitude, and financial capacity.

The graphical content is tailored to each client’s assigned risk profile, which ranges from 2 to 10. Crucially, the explanation is profile-based: for each risk profile, the bar chart illustrates the mean nor-



(a) Recommendation Interface for Non-Treated Users, translated from French



(b) Recommendation Interface for Treated Users, translated from French

malized importance of each determinant across all users assigned that same score. In other words, it shows what characterizes the "typical" individual in that risk category, offering users comparative insights into what the algorithm considers the average user within this risk class. Displayed "importance" reflects model weights and is not a causal attribution. This approach embeds the explanation within a broader reference frame and seeks to enhance interpretability by contextualizing the recommendation with feature-importance cues linked to the recommended risk profile.

The financial algorithm underlying the robo-advisor aggregates responses from a financial questionnaire to compute a global risk score. Each response contributes to the final score based on a predefined weighting system that reflects the relative importance of various attributes in assessing a client's investment profile. Although the scoring process is non-linear—such that small changes in inputs do not always produce proportional changes in the output—it is internally transparent: the structure and logic of the model are fully observable to the platform designers. In other words, the algorithm qualifies as a "white-box" model from a backend perspective. The display is a faithful transformation of the production scoring logic with no surrogate or post-hoc explainer used. This internal transparency ensures that the graphical representation used in the intervention remains an accurate reflection of the underlying

risk assessment process, allowing for a reliable interpretation of how various factors influence the final investment profile. However, this interpretability is not directly available to end users, who only see the final score without insight into how their responses shaped it. The graphical explanation provided in our intervention serves to bridge this gap by offering users an interpretable summary of the main factors influencing their assigned risk profile, thereby reflecting the logic of the underlying model in an accessible and meaningful way to prospective clients.

Our approach builds on established techniques in the XAI literature, particularly feature importance attribution methods that highlight which inputs drive model outputs. While our algorithm is not a black-box model requiring post-hoc methods like SHAP, the graphical explanation serves a similar function: to make the recommendation logic more transparent and intuitive. By delivering personalized, profile-based explanations grounded in real user profiles, our design aims to foster greater comprehension and trust in algorithmic recommendations, contributing to the broader literature on interpretable decision support in finance.

2.3.3 Outcome variables

We exploit a rich array of outcome measures, which we organize into two broad categories: portfolio choices, capturing how users made decisions once they receive a risk-score recommendation, and engagement outcomes, reflecting the intensity and nature of interaction with the “Tunnel” interface.

Portfolio choices focus on (i) the acceptance of the recommended risk profile; (ii) the deviation, which retains the sign of the deviation to distinguish upward from downward adjustments; and (iii) separate positive and negative deviations to gauge the intensity of upward versus downward modifications.

Engagement outcomes are inspired by the XAI and recommender-systems literatures and quantify user interactions with the platform: (i) the total number of attempts across all steps (whether or not a contract was ultimately signed); (ii) the number of opening events, defined as instances in which users clicked to access the modification interface; and (iii) the number of modification events, defined as instances in which users actively adjusted their risk profile within that interface.

Further details on the distribution of these outcome variables can be found in Table 2.A.2 in Appendix 2.6.2.

Portfolio Choices

Acceptance of score recommendation We define a binary variable $Accept_i$ that equals 1 if user i selects exactly the risk score recommended by the robo-advisor, and 0 otherwise. This outcome measures users’ willingness to follow the algorithmic suggestion.

Deviation measures from the recommended score We capture three related measures of how users adjust their chosen score relative to the robo-advisor’s recommendation.

First, the Deviation D_i preserves the sign of the difference—positive for upward adjustments and negative for downward adjustments—thereby indicating the deviation direction in risk-profile notches. Formally:

$$D_i = \text{ChosenRP}_i - \text{RecommendedRP}_i.$$

In this measure, when clients chose a score that is higher (lower) than the recommended score, the deviation is recorded using a positive (negative) value. This allows for documenting the direction of deviation from the recommended score, accounting for the sign of the deviation, if there is one.

Second and third, we subset signed deviations into separate measures of upward and downward modifications, respectively denoted PD_i and ND_i . Then, on the subsample with $D_i > 0$ we define

$$PD_i = D_i \quad (\text{for } D_i > 0),$$

and, on the subsample with $D_i < 0$ we define

$$ND_i = D_i \quad (\text{for } D_i < 0).$$

Engagement Outcomes

Total number of attempts We define $TotAttempts_i$ as the total count of times user i completed the Tunnel until Step 3, irrespective of whether a contract was ultimately signed. This aggregate metric captures overall engagement with the onboarding process.

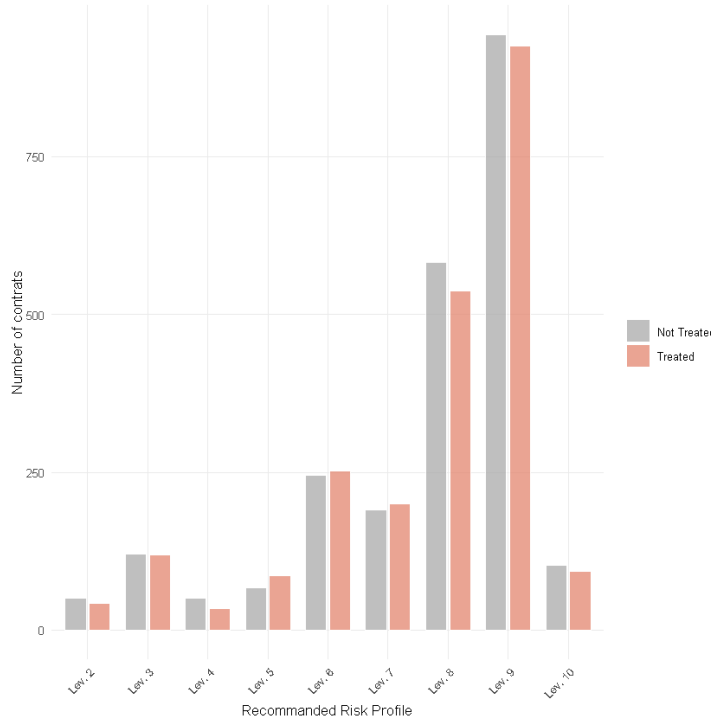
Number of opening events We define $OpenEvents_i$ as the number of instances in which user i clicked to access the modification interface. Each opening event reflects a user’s decision to review or reconsider the recommended risk profile.

Number of modification events We define $ModEvents_i$ as the number of instances in which user i actively adjusted their risk profile within the modification interface. This outcome measures the intensity of interaction with the recommendation system.

2.4 Results

2.4.1 Descriptive statistics

Figure 2.3: Distribution of Recommended Risk Profile by Experimental Condition



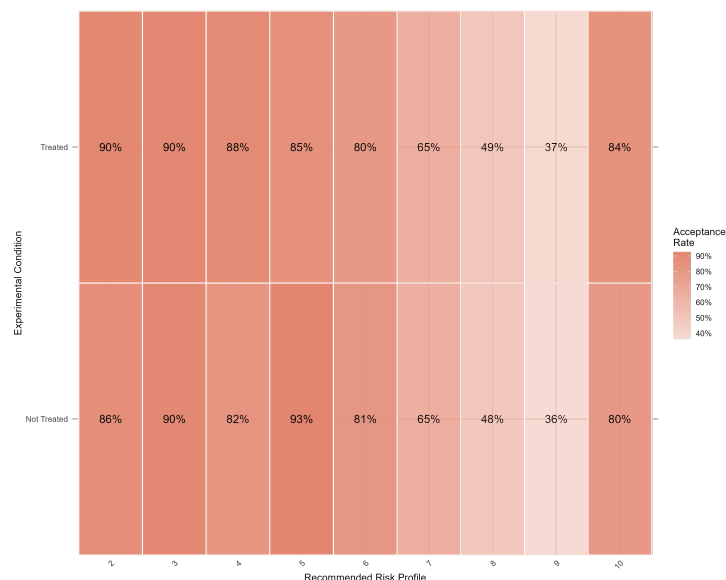
During the 105-day field experiment, 4,645 contracts³ were signed, corresponding to $N = 3,856$ unique users, with an average of 1.20 contracts per user ($SD = 0.57$). Of these, 2,354 belonged to the control group and 2,291 to the treatment group. The mean age of users was 38.00 years ($SD = 13.00$ years). Annual income was distributed as follows: 25.67% reported €30,000–45,000; 20.82% €60,000–100,000; 19.73% €45,000–60,000; 19.24 % < €30,000; 10.66% €100,000–150,000; and 3.86% > €150,000. Based on financial-knowledge items, 16.8% of users were classified as “*Non initié*” (uninitiated), 43.8% as “*Néophyte*” (beginner), and 39.5% as “*Sachant*” (knowledgeable). Finally, 86.87% of users were deemed novice in financial experience, while 13.12% were deemed experienced.

Baseline balance tests reveal no significant differences between the control and treatment groups in the distribution of recommended risk profiles (two-sided Mann–Whitney U test, p -value = 0.746; see Figure 2.3) nor in the distribution of device sessions (personal computer vs. phone; two-sided Mann–Whitney U test, p -value = 0.614; see Figure 2.A.5 in Appendix 2.6.2). Additionally, there were no differences in the distribution of user type — “first-time users”, who had not used the platform before the experiment,

³For confidentiality reasons, the total number of users who entered the Tunnel and the exact start and end dates of the experiment cannot be disclosed. Consequently, our analysis focuses on those who signed a contract and is conducted at the contract level.

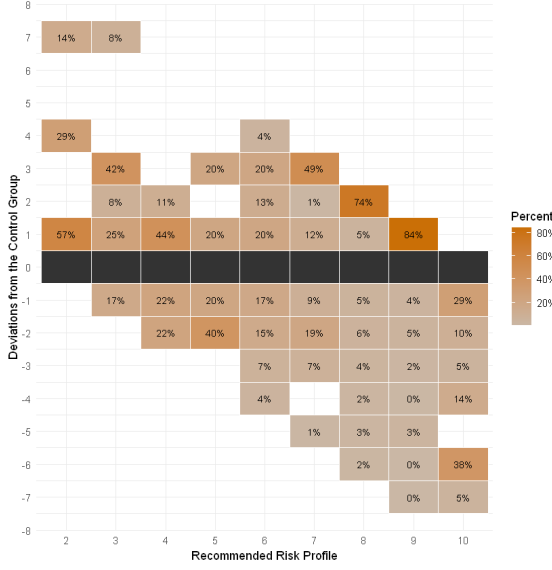
versus “returning users”, who had used the platform beforehand (two-sided Mann–Whitney U test, p -value = 0.742; see Figure 2.A.6 in Appendix 2.6.2).

Figure 2.4: Acceptance Rate per Recommended Risk Profile and Experimental Condition

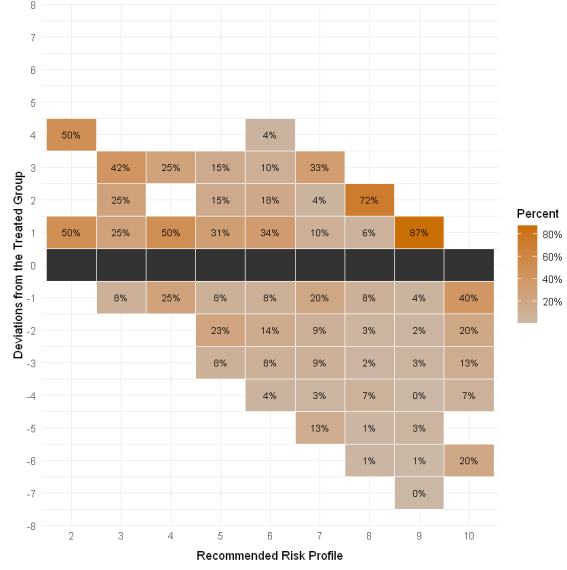


Descriptive statistics indicate that the overall acceptance rate does not differ significantly between the two experimental conditions: the control group exhibits a 54.63% acceptance rate, while the treatment group shows a 55.26% acceptance rate (two-sided Mann–Whitney U test, $p = 0.666$). Figure 2.4 displays the mean acceptance rate by risk profile for each group, revealing that both groups follow an almost identical pattern: acceptance rates range between approximately 80% and 90% for risk profiles 2 through 6 and profile 10, decline to below 70% for profile 7, drop below 50% for profile 8, and fall below 40% for profile 9. Notably, risk profiles 7–9 account for roughly 72% of our observations (see Table 2.A.3 in Appendix 2.6.2), underscoring the practical importance of these lower acceptance segments.

From the 2,093 observations in which users chose a different risk profile than recommended (1,645 positive deviations and 448 negative deviations), descriptive statistics reveal broadly similar deviation patterns across the two experimental conditions. The two heatmaps represent, for each recommended risk level, the percentage of users whose chosen profile lies at each deviation from the control group (Figure 2.5a) and the treated group (Figure 2.5b). In both graphs, deviations of +1 and +2 (i.e., selecting a risk profile one or two profiles higher than initially recommended) dominate at mid-range recommendations (profiles 3–7). For example, in the non-treated group, 42 percent of users recommended profile 3 actually selected profile 4 (+1 deviation), while 49 percent of users recommended profile 6 opted for profile 7 (+1). Similarly, in the treated group, 42 percent of profile 3 users and 33 percent of profile 6 users deviated upward by one notch. Upward deviations peak at recommended profile 8: 74 percent of non-treated users and 72 percent of treated users who deviated from it selected profile 10 (+2 deviation from 8).



(a) Deviations' Heatmap: Control Group



(b) Deviations' Heatmap: Treated Group

Notably, a substantial fraction of users recommended profiles 7 and 9 also deviated all the way to profile 10, suggesting they may have already had a predetermined preference for a fully equity-based portfolio. Negative deviations (selecting a lower risk profile than initially recommended) are relatively rare for mid-range recommendations but become more prevalent at the highest risk levels. For instance, among users recommended profile 10, 38 percent of treated users and 29 percent of non-treated users who deviated chose profile 4 (−6 deviation from 10), reflecting a strong downward shift for the riskiest recommendation. These patterns mirror the summary statistics in Table 2.A.4: both groups exhibit very similar weighted mean deviations for several recommended profiles (e.g., Profile 6 shows means of 0.367 in the non-treated group versus 0.400 in the treated group, and Profile 8 shows 0.907 versus 0.876), with modes typically at +1 for mid-range profiles.

2.4.2 Treatment Effect

Portfolio Choices

Table 2.2 reports the results of logistic regressions on the likelihood of users accepting the Robo-Advisor's recommended risk profile ($Accept_i$), pooling all experimental conditions (Models 1–4). In each specification, we include a comprehensive set of socio-demographic and investment-related controls (as detailed in the note). Model 1 shows that the treatment indicator (i.e., provision of an explanation) has a small, positive coefficient ($\beta = 0.014$), but is not statistically significant ($SE = 0.073$). Across Models 2–4, we introduce interaction terms between the treatment indicator and two user characteristics (detailed in Section 2.4.1): device session (Phone) and user type (First-Timers). In Model 2, the interaction between Treatment and Phone is negative ($\beta = -0.069$) but statistically insignificant ($SE = 0.148$),

Table 2.2: Explanations Effect on the Acceptance Rate

	<i>Dependent variable:</i>			
	<i>Accept_i</i>			
	(1)	(2)	(3)	(4)
Treated	0.014 (0.073)	0.049 (0.106)	0.016 (0.073)	-0.068 (0.118)
Phone	0.314*** (0.078)	0.348*** (0.108)	0.317*** (0.078)	0.319*** (0.078)
Treated × Phone		-0.069 (0.148)		
First-Timers			0.327*** (0.119)	0.259* (0.141)
Treated × First-Timers				0.139 (0.151)
Constant	2.458** (1.080)	2.437** (1.081)	2.119* (1.090)	2.161** (1.090)
Controls	Yes	Yes	Yes	Yes
Observations	4,645	4,645	4,645	4,645
Log Likelihood	-2,567.506	-2,567.380	-2,562.848	-2,562.366
Pseudo R ² (McFadden)	0.197	0.197	0.198	0.198

Note: Clustered robust standard errors by user ID are reported in parentheses. The regressions control for user age, type of investment project, investment horizon, initially declared investment amount, rank of tunnel run, estimated revenue, home-ownership status, financial experience, financial literacy and the initially recommended risk profile. . *p<0.1; **p<0.05; ***p<0.01

leaving the marginal effect of explanations on phone users negligible. In Model 3, the main effect of explanations remains insignificant and in Model 4, which adds the Treatment × First-Timers interaction, both treatment and the interaction terms are statistically indistinguishable from zero ($\beta = -0.068$, SE = 0.118 ; $\beta = 0.139$, SE = 0.151). Throughout all specifications, the Treatment coefficient does not attain significance, indicating that providing explanations does not meaningfully influence acceptance rates among users interacting with the same robo-advisor. By contrast, device session has a robust positive association with acceptance: users on mobile phones have approximately 37% higher odds to accept the recommendation than users on a computer (Model 1: $\beta = 0.314$, SE = 0.078, $p < 0.01$), an effect that persists across all specifications. Beyond device type, platform experience also matters. First-time users exhibit meaningfully higher acceptance than returning users (Model 3: $\beta = 0.327$, SE = 0.119, $p < 0.01$), corresponding to roughly 39% higher odds of accepting the recommendation; this attenuates to about 30% when the Treatment × First-Timers interaction is included (Model 4: $\beta = 0.259$, SE = 0.141, $p < 0.10$). Taken together, explanations do not measurably change acceptance, while device type and platform experience are the primary correlates of adherence.

Table 2.3: Explanations Effect on Deviations

	<i>Dependent variable:</i>			
	D_i			
	(1)	(2)	(3)	(4)
Treated	0.040 (0.078)	-0.029 (0.106)	0.039 (0.078)	0.093 (0.121)
Phone	-0.028 (0.079)	-0.098 (0.111)	-0.028 (0.079)	-0.030 (0.079)
Treated $\times$ Phone		0.144 (0.156)		
First-Timers			-0.028 (0.117)	0.020 (0.138)
Treated $\times$ First-Timers				-0.100 (0.157)
Constant	0.702 (1.099)	0.757 (1.107)	0.733 (1.102)	0.693 (1.098)
Controls	Yes	Yes	Yes	Yes
Observations	2,093	2,093	2,093	2,093
Adjusted R ²	0.221	0.221	0.221	0.221

Note: Clustered robust standard errors by user ID are reported in parentheses. The regressions control for user age, type of investment project, investment horizon, initially declared investment amount, rank of tunnel run, estimated revenue, home-ownership status, financial experience, financial literacy and the initially recommended risk profile. *p<0.1; **p<0.05; ***p<0.01

Table 2.3 presents OLS results for deviations (D_i) using again the 2,093 cases where users departed from the recommended profile. All four models control for demographics, investment characteristics, and initially recommended risk. Across all specifications, the treatment coefficient is small and not statistically different from zero. Likewise, being on a mobile device or a first-time user (and their interactions with treatment) shows no reliable impact on whether users shift toward higher or lower risk portfolios. The positive constant sign reflects the overall tendency to increase risk (1,645 positive vs. 448 negative deviations), but findings show that providing explanations does not shift that underlying direction.

Table 2.A.5 in Appendix 2.6.2 reports OLS estimates for positive deviation (PD_i), including only the 1,645 cases where users chose a higher risk profile than recommended. All four regressions control for the same socio-demographic, investment variables, and initially recommended risk. In Model 1, the explanation treatment has a small, negative coefficient ($\beta = -0.022$, SE = 0.018) that is not statistically significant, indicating no appreciable shift in upward deviations. Adding the Treatment $\times$ Phone interaction in Model 2 does not change this result ($\beta = -0.024$, SE = 0.027; interaction $\beta = 0.003$, SE =

0.040). Likewise, Model 3 and 4 also shows that treated users exhibit a slightly smaller upward deviation (3, $\beta = -0.022$, SE = 0.018; 4, $\beta = -0.024$, SE = 0.027) but without significance, and Model 4's Treatment $\times$ First-Timers interaction is effectively zero ($\beta = 0.003$, SE = 0.038). Across all specifications, no coefficient on treatment, device, or user type meaningfully alters the magnitude of upward shifts.

Table 2.4: Explanations Effect on Negative Deviations

	<i>Dependent variable:</i>			
	ND_i			
	(1)	(2)	(3)	(4)
Treated	-0.244 (0.150)	-0.438** (0.188)	-0.137 (0.148)	0.091 (0.241)
Phone	-0.059 (0.162)	-0.277 (0.223)	-0.049 (0.160)	-0.071 (0.159)
Treated $\times$ Phone		0.449 (0.324)		
First-Timers			0.787*** (0.214)	0.983*** (0.266)
Treated $\times$ First-Timers				-0.424 (0.304)
Constant	-1.488 (0.921)	-1.368 (0.906)	-3.098*** (1.051)	-3.207*** (1.052)
Controls	Yes	Yes	Yes	Yes
Observations	448	448	448	448
Adjusted R ²	0.134	0.136	0.156	0.158

Note: Clustered robust standard errors by user ID are reported in parentheses. The regressions control for user age, type of investment project, investment horizon, initially declared investment amount, rank of tunnel run, estimated revenue, home-ownership status, financial experience, financial literacy and the initially recommended risk profile.. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 2.4 focuses on the 448 instances where users moved to a lower-risk portfolio than recommended, estimating how explanations affect the magnitude of these downward shifts. In Model 1, the treatment coefficient is negative and non-significant (Model 1: $\beta = -0.244$, SE = 0.150). Adding the Treated $\times$ Phone term in Model 2 increases both the magnitude and significance of the treatment coefficient (Model 2: $\beta = -0.452$, SE = 0.189, $p < 0.05$), but the interaction itself is not significant, suggesting that treatment effects vary across devices. While there is no statistical difference between phone users, it seems that treated computer users exhibit larger downward moves than their untreated counterparts. However, once we account for platform experience, the treatment effect loses significance (Model 3: $\beta = -0.137$, n.s.; Model 4: $\beta = 0.091$, n.s.), whereas first-time users exhibit markedly smaller downward deviations (≈ 0.79 – 0.98 , $p < 0.01$). Complementary robustness checks adding the Treated $\times$ Phone $\times$ First-Timers interaction yield no reliable differential treatment effect by device or tenure (Appendix 2.6.2, Table 2.A.6). We

therefore interpret the desktop-specific amplification seen in simpler models as sensitive to specification; after accounting for user experience, there is no robust evidence that explanations increase the size of downward deviations.

In summary, transparency does not increase acceptance and does not affect deviations. The salient systematic differences are by device: desktop sessions accept less and deviate more than phone sessions. In simpler specifications, explanations are associated with larger downward moves among desktop users who already choose safer-than-recommended portfolios; however, this association disappears and becomes statistically insignificant once we account for user experience, and robustness checks with higher-order interactions do not reveal a stable differential effect by device or tenure. Taken together, the evidence points to behavior being shaped primarily by device context and platform experience, rather than by the explanation per se: adherence is higher on mobile and among first-time users; desktop users are more inclined to adjust, but explanations do not reliably push them further. Additionally, the overall deviation patterns, particularly the large shifts among those moving to a full-equity portfolio, suggest that many users may have already held a predetermined investment preference, rendering the robo-advisor's recommendation largely irrelevant regardless of treatment.

Engagement Outcomes

Across all three engagement measures—total attempts (Table 2.A.7, Appendix 2.6.2), “open” clicks to access the modification page (Table 2.A.8, Appendix 2.6.2), and actual modification events (Table 2.A.9, Appendix 2.6.2)—providing a profile-based graphical explanation had no impact on user behavior. In models predicting the total number of attempts, users who received an explanation neither reentered the tunnel process more nor fewer times than those in the control group, suggesting that explanations did not alter users' persistence or hesitancy during the onboarding process. Similarly, when examining how often participants clicked on the option to open the modification page, the only hint of a treatment effect was a small increase among treated desktop users when compared to untreated desktop users; however, this increase does not remain statistically significant once mobile usage and other covariates were taken into account, indicating no robust difference in exploratory behavior attributable to explanations. Finally, the count of actual modification events remained unchanged, implying that explanations did not spur users to explore other profiles by revising their choices more frequently. Engagement nonetheless differs systematically by device: phone sessions exhibit fewer openings of the modification option and fewer recorded modifications than desktop sessions (Tables 2.A.8–2.A.9; baseline coefficients around -0.34 for opens and -0.37 for modifications). This device pattern aligns with our main behavioral results: mobile users are more likely to accept the initial recommendation with minimal exploration, whereas desktop users explore more. Moreover, platform experience is a second, strong correlate of engagement.

First-time users are markedly more active across all three measures, with large positive coefficients for attempts (≈ 0.73 - 0.73), opens (≈ 0.84 - 0.87), and modifications (≈ 0.75 - 0.85), all significant at $p < 0.01$ (Tables 2.A.7–2.A.9). Collectively, these findings demonstrate that, contrary to expectations from human–computer interaction and XAI research, displaying how the robo-advisor positions each user relative to the “average” risk profile neither increased nor decreased engagement. In other words, revealing prototypical characteristics for each risk class did not lead participants to explore more options, revisit the onboarding process more often, or investigate alternate portfolios. This suggests that engagement on this platform is shaped by factors beyond algorithmic transparency, such as the type of device used and user experience.

Treatment Heterogeneity

To this point, our analysis has focused on estimating the average treatment effect (ATE) of providing peer-based graphical explanations versus no explanation. Formally, the ATE is defined as

$$\tau = E[Y_i | D_i = 1] - E[Y_i | D_i = 0],$$

where Y_i denotes the outcome of interest (e.g., behavioral measures, engagement measures) and $D_i \in \{0, 1\}$ indicates whether user i received the explanation. While the ATE captures the overall effect across our entire sample, it masks any systematic variation in treatment response that may depend on observable characteristics.

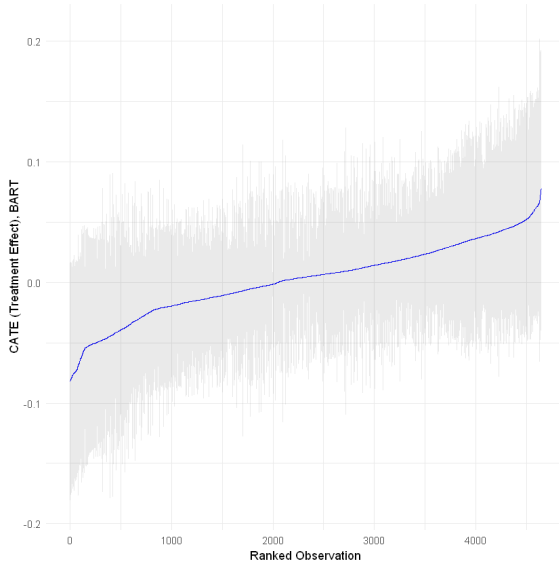
To uncover such heterogeneity, we turn to recent advances in causal machine learning—specifically, the Conditional Average Treatment Effect (CATE) framework. The CATE for a user with covariate vector $X_i = x$ is defined as

$$\tau(x) = E[Y_i | D_i = 1, X_i = x] - E[Y_i | D_i = 0, X_i = x].$$

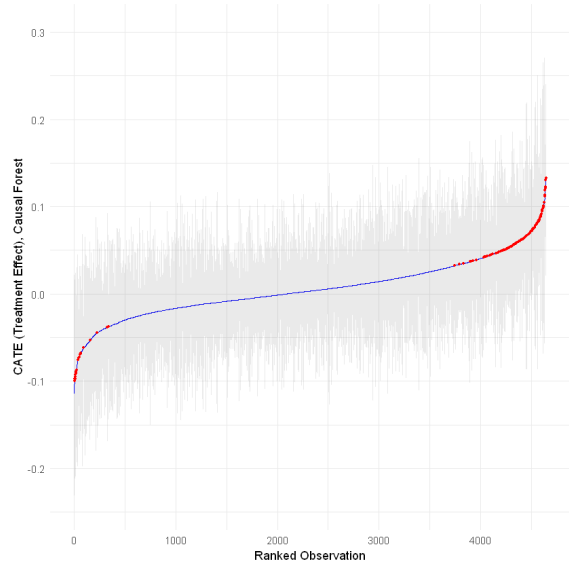
By estimating $\tau(x)$, we can identify subpopulations for whom explanations are particularly effective (or counterproductive). Following the approach outlined in Jacob (2021), we employ two flexible, non-parametric estimators well suited for high-dimensional covariate sets: Causal Random Forests (CRF) and Causal Bayesian Additive Regression Trees (BART). These methods combine ensemble learning with honest splitting rules to yield asymptotically unbiased estimates of $\tau(x)$ under standard unconfoundedness assumptions. In the sections that follow, we briefly describe each algorithm’s main features and then apply them to our experimental data, allowing us to explore whether treatment effects vary systematically by user demographics, device type, or prior financial experience.

For the present analysis, we limit ourselves to two primary outcomes: (i) acceptance of the Robo-

Advisor’s recommendation and (ii) the signed deviation from the recommended risk profile. Causal Random Forests and Bayesian Additive Regression Trees both draw on Jacob (2021) to flexibly estimate $\tau(x)$ without imposing rigid functional forms. As Jacob explains, Causal Random Forests grow an ensemble of “honest” trees—each splitting on a random subsample to form leaves with similar covariates, then using held-out observations to estimate within-leaf treatment and control means—so that averaging across trees yields unbiased, heterogeneity-aware estimates of $\tau(x)$. In contrast, BART fits the response surface as a sum of many small trees under regularizing priors, using MCMC to sample from the posterior of individual treatment effects; this approach naturally captures complex interactions and provides credible intervals for $\tau(x)$. Both methods rely on sample-splitting or honest estimation to avoid overfitting and, under unconfoundedness, recover the CATE for acceptance and relative deviation.



(a) Causal BART, Acceptance of the Recommendation



(b) Causal Forest, Acceptance of the Recommendation

Table 2.5: CATE results for the Acceptance of the Recommendation

Method	20% least	ATE	20% most
Causal BART	-0.0381	0.0023	0.0429
Causal Forest	-0.0368	0.0059	0.0532

Figure 2.6b confirms that Causal-BART yields a tightly centered CATE distribution, with no individual effects whose 95% credible intervals exclude zero, whereas Causal-Forest (Figure 2.6a) uncovers 143 users (out of 4,645) with credibly nonzero treatment effects (red dots). Table 2.5 presents the key heterogeneity summary, namely, the estimated CATE at the 20% least affected, the overall ATE, and the 20% most affected, for both Causal BART and Causal Forest. For the bottom quintile, BART yields a CATE of -0.0380 , whereas the Forest gives -0.0368 , demonstrating almost identical minimal

downward shifts. The ATEs are likewise nearly zero (0.0023 for BART; 0.0059 for CF). In the upper quintile, however, the two methods diverge: BART’s top-20% average is 0.0429, while the Forest’s is a higher 0.0532. This larger positive tail under Causal Forest reflects its tendency to predict more extreme effects at the distributional margins. Overall, BART’s CATEs span only -0.038 to $+0.043$, compared to -0.037 to $+0.053$ for the Forest, both centered closely on zero, indicating that, although a small subset of users shows credibly nonzero shifts in acceptance, the vast majority experience no meaningful change in compliance under either estimator.

Table 2.6: Characteristics of Users by CATE Group, Acceptance of the Recommendation

Variable	No Significance (n=4,502)	Positive (n=119)	Negative (n=24)	p-value
<i>Proportions</i>				
Low patrimony (<30 k)	0.1966	0.0924	0.2917	0.0087**
Project Type: Savings	0.8827	0.9664	0.9167	0.0164*

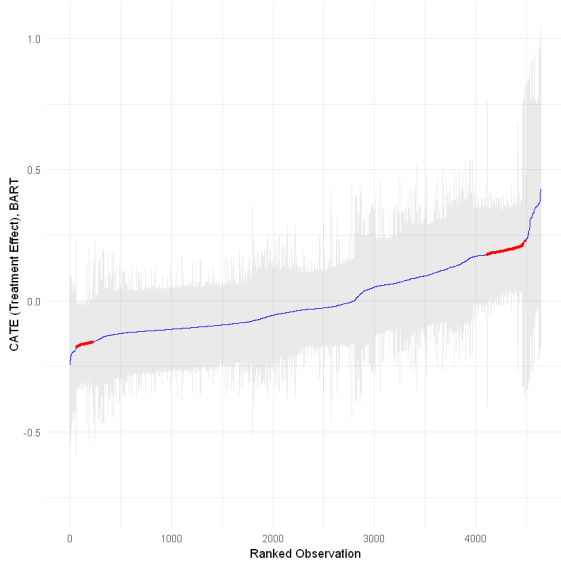
Note: Proportions are means of 0/1 indicators. p-values from one-way ANOVA across the three groups. *p<0.05; **p<0.01.

We find that only two user characteristics meaningfully influence the effect of peer-based explanations on acceptance (Table 2.6). First, investors in the lowest wealth bracket ($< \text{€}30\text{k}$) display systematically different responses: only 9.2% of those with a positive CATE fall into this category, versus 19.7% in the no-effect group, while 29.2% of the small Negative-CATE subgroup occupy the low-wealth segment. Second, a savings-oriented investment project objective robustly predicts positive responsiveness: 96.6% of the Positive-CATE users report a savings goal, compared to 88.3% in the Non Significant group ($p < 0.01$). In plain terms, explanations appeal most to relatively wealthier, savings-focused investors and can slightly induce more conservative adjustments among poorer clients. All other covariates—age, device type, tenure, baseline risk recommendation, and so on, exhibit no credible differences across CATE segments.

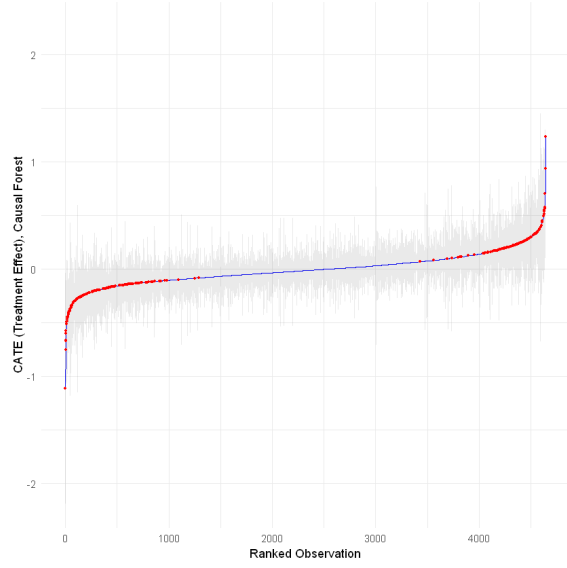
Table 2.7: CATE results for the Deviation

Method	20% least	ATE	20% most
Causal BART	-0.1368	0.0003	0.1979
Causal Forest	-0.1938	-0.005	0.2052

Figure 2.7a shows that Causal BART’s CATEs for signed deviation remain tightly concentrated around zero; only a relatively small subset of users exhibit credibly non-zero effects, whereas Causal Forest (Figure 2.7b) continues to flag somewhat more extreme effects in both tails. Table 2.7 reports the key quantile summaries of the CATE distribution for signed relative deviation under both estimators. Causal BART yields a 20% least-affected CATE of about -0.14 (roughly -13.7 pp), an overall mean effect (ATE) that is essentially zero (≈ 0.0003 , or $+0.03$ pp), and a 20% most-affected CATE of around



(a) Causal BART, Deviation



(b) Causal Forest, Deviation

+0.20 (+19.8 pp). Causal Forest produces a very similar central tendency ($ATE \approx -0.005$, i.e. about -0.5 pp) but somewhat more pronounced tails, with an average CATE of about -0.19 (-19.4 pp) in the lower quintile and $+0.21$ ($+20.5$ pp) in the upper quintile. In plain terms, while a minority of investors tighten or loosen their deviation by on the order of 15–20 percentage points, the bulk of the sample remains very close to zero, and both methods agree on an average effect that is effectively negligible.

Table 2.8: User Characteristics by CATE Group and Method (Deviation)

Variable	BART				Causal Forest			
	NoSig	Pos	Neg	p(BART)	NoSig	Pos	Neg	p(CF)
<i>n</i>	4,347	210	88	—	4,454	53	138	—
<i>Proportions</i>								
Financial Knowledge: Beginner	0.445	0.409	0.329	0.063*	—	—	—	—
Financial Knowledge: Knowledgeable	0.393	0.433	0.500	0.068*	—	—	—	—
Investment Horizon	10.806	11.685	11.579	0.065*	—	—	—	—
Project type: Savings	0.883	0.933	0.863	0.068***	0.885	0.940	0.821	< 0.01***
Project type: Retirement	—	—	—	—	0.044	0.006	0.076	< 0.01***
<i>Means</i>								
User age (years)	—	—	—	—	21.54	17.90	23.22	< 0.01***
Recommended risk profile	6.615	6.985	6.909	< 0.01***	—	—	—	—

Note: “NoSig,” “Pos,” and “Neg” denote non-significant, positive, and negative CATE estimates, respectively; proportions for binary indicators, means for continuous variables. p-values from one-way ANOVA across the three CATE groups, by method. * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table 2.8 shows that, among the observed covariates, the stated project type remains the clearest predictor of how much users change their signed deviation in response to peer-based explanations. Under both BART and Causal Forest, users in the Positive-CATE group are more likely than the NoSig group to report a “savings” objective (around 93–94% versus roughly 88–89%), whereas those in the Negative-CATE group are less likely to do so (about 86% under BART and 82% under Causal Forest). In the Forest estimates, this pattern is mirrored on the “retirement” margin: only about 0.6% of the Positive-

CATE group, compared with roughly 4% of the no-effect group and 7–8% of the Negative-CATE group, report retirement as their goal, and the corresponding p-values for the savings and retirement indicators are both below 1%. Other characteristics play a more muted role. In the BART results, investors with negative CATEs tend to be somewhat less likely to classify themselves as beginners and more likely to describe themselves as knowledgeable, have slightly longer investment horizons, and receive marginally higher recommended risk profiles than those with no significant effect, but these differences are modest in magnitude and only weakly significant (p-values around 0.06–0.07). In the Forest results, age emerges as the main additional discriminator: the Positive-CATE group is on average younger (about 18 years) than the no-effect group (around 21.5 years), while the Negative-CATE group is older (over 23 years), consistent with retirement-oriented investors being more inclined to reduce their deviation. Overall, it is primarily the investor’s goal—savings versus retirement—and, to a lesser extent, age and self-reported financial sophistication that differentiate who adjusts their deviation, whereas other demographics and financial attributes show only modest or method-specific splits across CATE segments.

In sum, our analyses paint a consistent picture: peer-based graphical explanations have essentially no impact on average acceptance or deviation, with ATEs close to zero across both Causal BART and Causal Forest. Treatment heterogeneity is limited and concentrated in small subgroups. For acceptance, two user characteristics stand out: investors in the lowest wealth bracket are somewhat more likely to exhibit negative CATEs, whereas relatively wealthier users with a savings-oriented project are more likely to display positive CATEs. For deviations, the investor’s stated objective—savings versus retirement—emerges as the most robust moderator under both estimators: retirement-oriented users tend to reduce their deviation from the recommendation, while savings-focused users tend to increase it, with other demographic and financial attributes showing at most weak or method-specific differences across CATE segments. Taken together, these results suggest that, for this platform and population, algorithmic transparency in the form of peer-based explanations does not meaningfully shift the average user’s behavior and only modestly influences the choices of a small, well-defined subset of investors.

2.5 Conclusion

This paper uses a randomized controlled trial embedded in the interface of a leading French robo-advisor to test whether profile-based graphical explanations influence portfolio choices. Two facts stand out. First, contrary to prevailing expectations in the XAI and behavioral finance literatures, explanations neither increase adherence to the recommended risk score nor reduce the magnitude of deviations when they opt to choose a different profile. Additionally, engagement metrics, such as the number of tunnel completions, interface openings, or actual profile modifications, remain statistically indistinguishable

across treatment and control groups. Second, behavior varies systematically by context: relative to phone sessions, users on desktop accept less and explore more, and first-time users accept more and engage more than returning users. In some simpler specifications we observe larger downward moves among desktop users who already choose a safer-than-recommended portfolio. However, this pattern disappears and becomes statistically indistinguishable from zero once we account for user experience, and additional robustness checks do not reveal a stable differential treatment effect by device or tenure.

These null results carry important implications for both theory and practice. First, they suggest that in high-stakes financial decisions, where many investors appear to hold ex-ante portfolio preferences, algorithmic explanations may be insufficient to overcome deeply held convictions or “algorithm aversion.” Second, they highlight the need to reconsider the design of explainability interventions: peer-based benchmarks alone may not resonate with end-users unless coupled with interactive, personalized, or narrative components that address individual concerns more directly. Third, from a policy standpoint, our findings caution regulators and platform designers against assuming that transparency mandates will automatically translate into greater adoption or trust without complementary measures.

If explanations matter, they appear to do so through context rather than on average. Desktop interactions are associated with more deliberation and adjustment; mobile interactions with higher baseline adherence and fewer adjustments. But once user experience is taken into account, we do not find robust evidence that explanations systematically push desktop users further toward lower risk. The design take-away is therefore not “add explanations everywhere,” but “tailor presentation to context and experience.”

For design and policy, the implication is to favor context-contingent transparency over one-size-fits-all. Platforms aiming to preserve adherence at the acceptance step could test lighter or deferred explanations on desktop (e.g., progressive disclosure, post-confirmation details, or optional “learn more”), evaluate alternative benchmark framings, and segment by tenure. Conversely, mobile flows may not warrant additional surface area for explanations at the decision point. More broadly, transparency alone may shift adjustments rather than the decision to accept itself.

Our setting bounds inference. We study a white-box, peer-based explanation during onboarding for life-insurance wrappers, observe short-run choices (not funding, rebalancing, or performance), and evaluate one market and design. Other explanation styles (e.g., counterfactuals, conversational agents) or products may behave differently, and longer horizons may reveal delayed trust effects.

Overall, the evidence maps boundary conditions for XAI in digital financial advice: no average improvement in acceptance, behavior shaped primarily by device context and platform experience, and only weak, specification-sensitive signs that explanations could magnify downward moves on desktop. While algorithmic transparency remains an essential component of accountable FinTech, effective explainability will likely come from adaptive, device and experience-focused designs when adherence is desired; richer

and interactive when exploration is the goal rather than uniform displays of feature importance. Future work should test such adaptive frameworks across investor segments and over extended horizons.

2.6 Appendix

2.6.1 Experimental Design Appendices

Figure 2.A.1: Graphical explanation for the Risk Profile 10

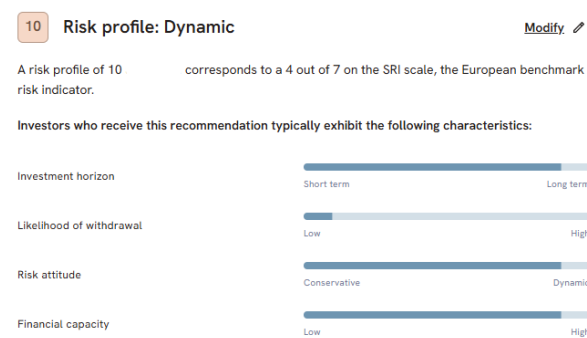


Figure 2.A.2: Graphical explanation for the Risk Profile 6

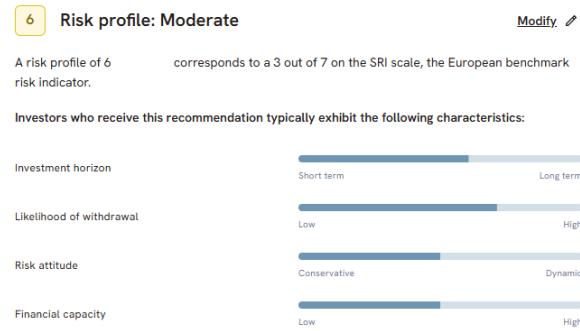


Figure 2.A.3: Graphical explanation for the Risk Profile 3

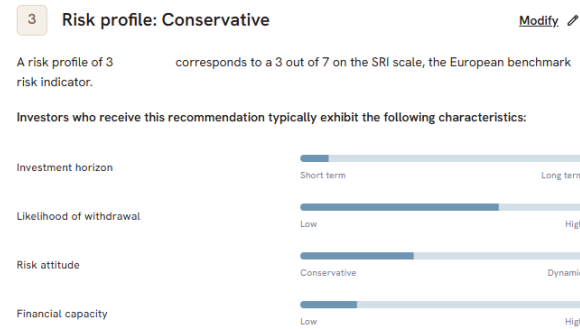


Figure 2.A.4: Modification Page

Modification of your allocation ✕

Your allocation can be modified at any time from your investor portal.

Responsible investment   

Selected risk profile: Profile 9

- ☐ Suravenir Opportunités 2 Euro Fund
- ☐ HSBC Clic Euro 85
- ☐ SC Real Estate
- ☐ Transparency III
- ☐ 2 Profile 2
- ☐ 3 Profile 3
- ☐ 4 Profile 4
- ☐ 5 Profile 5
- ☐ 6 Profile 6
- ☐ 7 Profile 7
- ☐ 8 Profile 8
- ☒ 9 Profile 9 Recommended for you
- ☐ 10 Profile 10

You want to take on significant risk with the prospect of high returns and you have a long term investment horizon (>8 years). Profile 9 is therefore perfectly suited to your project.

Confirm

2.6.2 Additional Results

Table 2.A.1: Balance tests between treatment and control groups (non-parametric)

Variable	<i>p</i> -value	Bonferroni <i>p</i>
<i>Numeric / ordinal covariates (Mann–Whitney)</i>		
User age	0.119	0.476
Initially declared investment amount	0.600	1.000
Rank of tunnel run	0.853	0.968
Initially recommended risk profile	0.749	1.000
<i>Categorical covariates (χ^2 with Fisher fallback)</i>		
Financial experience	0.625	1.000
Estimated Revenue	0.767	1.000
Type of investment project	0.571	1.000
Home-ownership status	0.872	1.000
Financial literacy	0.978	1.000
Investment Horizon	0.761	1.000

Notes: Two-sided tests. Numeric/ordinal covariates tested with Mann–Whitney (Wilcoxon rank-sum). Categorical covariates tested with Pearson χ^2 ; Fisher’s exact used when expected counts are small. Bonferroni adjustment applied within each family (numeric vs. categorical). Values rounded to three decimals.

Figure 2.A.5: Distribution of Devices Used by Experimental Condition

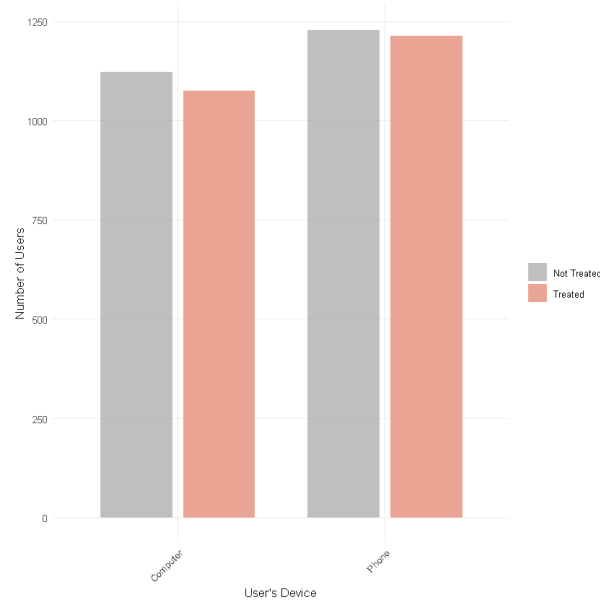


Figure 2.A.6: Distribution of Users' Type by Experimental Condition

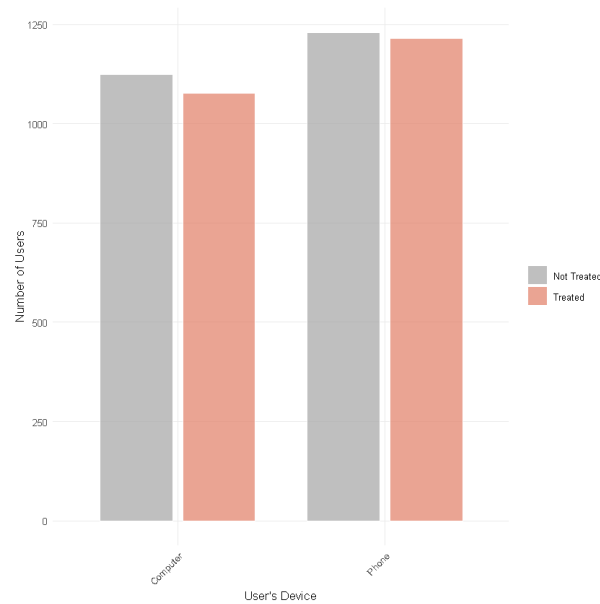


Table 2.A.2: Summary Statistics for Behavioral and Engagement Outcomes

Outcome	Q1	Mean	Median	Q3	SD	N
Behavioral Outcomes						
Acceptance ($Accept_i$)	0.00	0.55	1.00	1.00	0.49	4,645
Deviation (D_i)	1.00	0.52	1.00	2.00	1.91	2,093
Positive deviation (PD_i)	1.00	1.39	1.00	2.00	0.64	1,645
Negative deviation (ND_i)	-4.00	-2.65	-2.00	-1.00	0.19	448
Engagement Outcomes						
Total attempts ($TotAttempts_i$)	1.00	1.48	1.00	2.00	0.99	3,856
Opening events ($OpenEvents_i$)	0.00	1.14	0.00	1.00	3.69	3,856
Modification events ($ModEvents_i$)	0.00	0.71	0.00	0.00	2.83	3,856

Table 2.A.3: Number of observations per recommended risk level by experimental condition and in total.

Risk Profile	Not Treated	Treated	Total	Overall (%)
Profile 2	51	42	93	2.0
Profile 3	121	119	240	5.2
Profile 4	51	34	85	1.8
Profile 5	67	87	154	3.3
Profile 6	245	252	497	10.7
Profile 7	191	200	391	8.4
Profile 8	582	538	1120	24.1
Profile 9	943	926	1869	40.2
Profile 10	103	93	196	4.2

Table 2.A.4: Summary Statistics of Deviation by Recommended Risk Level and Experimental Condition

Risk Profile	Non-treated			Treated		
	Total n	Weighted Mean	Mode	Total n	Weighted Mean	Mode
Profile 2	7	2.7143	1	4	2.5000	1
Profile 3	12	2.0833	3	12	1.9167	3
Profile 4	9	0.0000	1	4	1.0000	1
Profile 5	5	-0.2000	-2	13	0.3077	1
Profile 6	46	0.3696	1	50	0.4000	1
Profile 7	67	0.8507	3	70	-0.2143	3
Profile 8	300	0.9067	2	275	0.8764	2
Profile 9	601	0.4093	1	582	0.5206	1
Profile 10	21	-3.8095	-6	15	-2.6667	-1

Table 2.A.5: Explanations Effect on Positive Deviations

	<i>Dependent variable:</i>			
	PD_i			
	(1)	(2)	(3)	(4)
Treated	−0.022 (0.018)	−0.024 (0.027)	−0.022 (0.018)	−0.024 (0.027)
Phone	0.007 (0.019)	0.005 (0.031)	0.007 (0.019)	0.007 (0.019)
Treated × Phone		0.003 (0.040)		
First-Timers			−0.007 (0.027)	−0.008 (0.035)
Treated × First-Timers				0.003 (0.038)
Constant	2.136* (1.174)	2.137* (1.176)	2.143* (1.175)	2.144* (1.171)
Controls	Yes	Yes	Yes	Yes
Observations	1,645	1,645	1,645	1,645
Adjusted R ²	0.668	0.668	0.668	0.668

Note: Clustered robust standard errors by user ID are reported in parentheses. The regressions control for user age, type of investment project, investment horizon, initially declared investment amount, rank of tunnel run, estimated revenue, home-ownership status, financial experience, financial literacy and the initially recommended risk profile.

*p<0.1; **p<0.05; ***p<0.01

Table 2.A.6: Explanations Effect on Negative Deviations (Cont.)

	<i>Dependent variable:</i>		
	ND_i		
	(1)	(2)	(3)
Treated	−0.123 (0.270)	−0.131 (0.268)	−0.265 (0.317)
Phone	−0.302 (0.218)	−0.158 (0.288)	−0.295 (0.345)
Treated × Phone	0.480 (0.319)	0.471 (0.320)	0.751 (0.493)
First-Timers	0.990*** (0.263)	1.096*** (0.289)	0.988*** (0.314)
Treated × First-Timers	−0.406 (0.303)	−0.406 (0.303)	−0.176 (0.399)
Phone × First-Timers		−0.251 (0.288)	−0.004 (0.425)
Treated × Phone × First-Timers			−0.510 (0.595)
Constant	−3.105*** (1.040)	−3.260*** (1.053)	−3.176*** (1.066)
Controls	Yes	Yes	Yes
Observations	448	448	448
Adjusted R ²	0.160	0.159	0.158

Note: Clustered robust standard errors by user ID are reported in parentheses. The regressions control for user age, type of investment project, investment horizon, initially declared investment amount, rank of tunnel run, estimated revenue, home-ownership status, financial experience, financial literacy and the initially recommended risk profile.
 *p<0.1; **p<0.05; ***p<0.01

Table 2.A.7: Explanations Effect on the Total Number of Attempts

	<i>Dependent variable:</i>			
	<i>TotAttempts_i</i>			
	(1)	(2)	(3)	(4)
Treated	−0.007 (0.022)	0.009 (0.030)	−0.001 (0.020)	0.002 (0.034)
Phone	0.012 (0.023)	0.027 (0.031)	0.013 (0.021)	0.013 (0.021)
Treated × Phone		−0.029 (0.042)		
First-Timers			0.726*** (0.031)	0.728*** (0.038)
Treated × First-Timers				−0.004 (0.041)
Constant	0.149 (0.143)	0.141 (0.145)	−0.577*** (0.137)	−0.578*** (0.137)
Controls	Yes	Yes	Yes	Yes
Observations	3,856	3,856	3,856	3,856
Log Likelihood	−5,042.480	−5,042.332	−4,912.462	−4,912.459
Pseudo-R ² (McFadden)	0.030	0.030	0.055	0.055

Note: Robust standard errors are reported in parentheses. The regressions control for user age, type of investment project, investment horizon, initially declared investment amount, rank of tunnel run, estimated revenue, home-ownership status, financial experience, financial literacy, and the initially recommended risk profile. *p<0.1; **p<0.05; ***p<0.01

Table 2.A.8: Explanations Effect on the Total Number of Open Events

	<i>Dependent variable:</i>			
	<i>OpenEvents_i</i>			
	(1)	(2)	(3)	(4)
Treated	0.065 (0.101)	0.212* (0.123)	0.071 (0.101)	0.102 (0.171)
Phone	−0.340*** (0.094)	−0.171 (0.139)	−0.337*** (0.093)	−0.338*** (0.092)
Treated × Phone		−0.336* (0.195)		
First-Timers			0.839*** (0.139)	0.867*** (0.177)
Treated × First-Timers				−0.056 (0.209)
Constant	−1.084* (0.609)	−1.170* (0.609)	−1.970*** (0.658)	−1.989*** (0.664)
Controls	Yes	Yes	Yes	Yes
Observations	3,856	3,856	3,856	3,856
Log Likelihood	−7,918.582	−7,904.173	−7,779.042	−7,778.649
Pseudo-R ²	0.141	0.142	0.156	0.156

Note: Robust standard errors are reported in parentheses. The regressions control for user age, type of investment project, investment horizon, initially declared investment amount, rank of tunnel run, estimated revenue, home-ownership status, financial experience, financial literacy, and the initially recommended risk profile. *p<0.1; **p<0.05; ***p<0.01

Table 2.A.9: Explanations Effect on the Total Number of Modification Events

	<i>Dependent variable:</i>			
	<i>ModEvents_i</i>			
	(1)	(2)	(3)	(4)
Treated	0.051 (0.117)	0.202 (0.142)	0.056 (0.117)	0.171 (0.189)
Phone	-0.366*** (0.108)	-0.191 (0.163)	-0.362*** (0.108)	-0.366*** (0.107)
Treated × Phone		-0.349 (0.224)		
First-Timers			0.747*** (0.174)	0.850*** (0.215)
Treated × First-Timers				-0.211 (0.234)
Constant	-1.689** (0.680)	-1.777*** (0.683)	-2.484*** (0.728)	-2.550*** (0.734)
Controls	Yes	Yes	Yes	Yes
Observations	3,856	3,856	3,856	3,856
Log Likelihood	-5,676.116	-5,666.519	-5,607.795	-5,604.267
Pseudo-R ²	0.152	0.154	0.162	0.163

Note: Robust standard errors are reported in parentheses. The regressions control for user age, type of investment project, investment horizon, initially declared investment amount, rank of tunnel run, estimated revenue, home-ownership status, financial experience, financial literacy, and the initially recommended risk profile. *p<0.1; **p<0.05; ***p<0.01

Chapter 3

Chapter 3

Algo-Rhythm Unplugged: Effects of explaining algorithmic recommendations on music discovery

Co-written with Julien M'Barki

3.1 Introduction

Over the past few decades, the music industry has undergone significant changes, engendering a profound impact on the dynamics of music consumption. Pioneered by platforms such as Spotify, Deezer, or Apple Music, this technological evolution has made it possible for listeners to access an expansive repertoire of musical compositions spanning diverse genres, artists, and historical periods. The sheer volume of content available on music streaming services has concurrently stimulated the need for widespread content recommendations, aiming to reduce users' search efforts (J. Y. Bakos, 1997). Traditionally, the literature identifies four primary modes of recommendation: editorial, self-perpetuated, contributing, and personalized (Farchy et al., 2017). Notably, personalized recommendation on streaming platforms predominantly takes the shape of algorithmic recommendations derived from user preferences, historical listening patterns, and genre affinities, commonly referred to as Music Recommender Systems (MRS) in the literature. Consequently, the transition from traditional methods of music consumption to on-demand streaming has not only revolutionized distribution models but has also intricately reshaped the landscape of music discovery, with a significant role played by sophisticated MRSs integrated into streaming platforms (Moreau & Wikström, 2024).

In other sectors of the economy, sophisticated digital technology, big data analytics, and continuous advancements in Artificial Intelligence (AI) technologies, such as machine learning (ML) and natural language processing (NLP), have also empowered individuals and organizations to delegate decision-making to algorithms. This phenomenon is observable in a wide range of domains, ranging from objective tasks to those requiring subjective evaluation. In healthcare, algorithms are integral to imaging, diagnosis, surgery, personalized medicine, or detecting infected persons (Castelluccia & Le Métayer, 2019; Haque et al., 2021). Robo-advisors leverage algorithms for investment decisions (L. Zhang et al., 2021) while managers employ them to forecast product demand and make informed hiring decisions (Diab et al., 2011; Kawaguchi, 2021). In courts, AI systems assess the likelihood of recidivism and the likelihood of flight of those awaiting trial, or of offenders in bail and parole procedures (Završnik, 2020). Furthermore, marketers utilize algorithms in areas such as research, strategy formulation, channel management, performance evaluation, and online advertising (Eslami et al., 2018; De Bruyn et al., 2020; Huang & Rust, 2021).

The rising use of algorithms to delegate tasks in diverse facets of our lives renewed discussions on autonomy, freedom, and the intrinsic value of decision rights (Bartling et al., 2014). As these algorithms play an increasingly prominent role, exploring their impact on decision rights becomes essential. The exploration of this effect is not confined solely to philosophical considerations; it extends to economic decisions, where the delegation of tasks to algorithms can yield far-reaching consequences. End-users of algorithmic recommendation systems face an agency problem due to information asymmetry and misaligned incentives between the user and the algorithm provider. While users (the principals) rely on these systems to deliver personalized and relevant suggestions, they often lack insight into how these algorithms operate and cannot verify whether the recommendations align with their true preferences. The platform or provider (the agent) may have different objectives, such as maximizing engagement or promoting specific content for commercial purposes, which can lead to “biased” or sub-optimal recommendations. Additionally, users have limited control over the algorithm’s decision-making process and may not be aware of how their data is used to influence these suggestions. This lack of transparency and control compounds the agency problem, as users cannot be certain that the system is acting in their best interest, potentially leading to inefficiencies in their decision-making and consumption behavior.

In response to these challenges, public policies, specifically at the European level, have developed a framework to improve transparency and interpretability in algorithmic decision-making. These regulations, such as the Digital Service Act¹ (DSA) or the AI Act,² have introduced an obligation to provide explanations in recommender systems. Specifically, the DSA requires the companies “to set out in plain

¹Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 (Digital Services Act).

²Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 (Artificial Intelligence Act).

and intelligible language the main parameters used in their recommender systems, as well as any options for the recipients of the service to modify or influence those main parameters. The main parameters shall explain why certain information is suggested to the recipient of the service. They shall include, at least, the criteria which are most significant in determining the information suggested to the recipient of the service and the reasons for the relative importance of those parameters.” In addition, the AI Act requires providers to “ensure that AI systems intended to interact directly with persons are designed and developed in such a way that individuals are informed that they are interacting with an AI system.” Music streaming services are subject to these regulations and must comply with them by introducing explanation systems to their MRSs. This is the case, for example, with Spotify, which recently introduced an explanation text of its personalized recommendation playlists.

Using a laboratory experiment, we study music discovery behaviors. Our design consists of a listening task with a duration of 20 minutes. The participants receive a set of recommendations generated by a simplified MRS based on a music preferences questionnaire. Our main measurement is the relative listening time of each track and a comparison between three groups: a control group that receives no explanations of the MRS generation process and recommendations (*Algorithm*); a control group that receives no explanations of the MRS generation process and recommendations based on the reversal of their preferences (*Anti-Algorithm*); and finally, a treatment group that receives explanations on the MRS generation process (*Explanations*). Rather than asking whether explanations make recommender systems more effective—a question extensively studied in the literature—our experiment focuses on whether explanations induce a shift in users’ discovery behaviors. In other words, we investigate whether sharing explanations about the recommendation process influences the way individuals explore, allocate attention, and diversify their listening during music discovery. Our research question is therefore: Do explanations of music recommender systems induce a shift in users’ discovery behaviors?

Although our experiment focuses on a highly specific setup—music discovery through a simplified MRS—the insights derived from this study have broader implications for other algorithmic decision-making environments. Music discovery provides an ideal testbed to study this issue: it is a low-stakes, repeated, and measurable environment where we can isolate how explanations affect user behavior, without the confounding risks of high-stakes domains. Similar agency problems arise in e-commerce, job matching, and even algorithmic pricing, where users must decide whether to trust and act upon recommendations or seek alternatives. Furthermore, our investigation into the impact of explanations aligns with broader discussions regarding European regulations, public policies, and Explainable AI (XAI) literature regarding transparency, trust, and user autonomy. By analyzing how explanations influence belief formation and decision-making in music discovery, we contribute to a general understanding of how algorithmic explainability can affect user behavior across diverse applications, from financial robo-advising to personalized

healthcare recommendations.

This paper is composed of 6 sections after this introduction: Section 3.2 presents the state-of-the-art regarding XAI and its articulation with MRSs. Section 3.3 presents the microeconomic theoretical framework we rely on to characterize the music discovery behaviors we measure. Section 3.4 presents the experimental design in detail with respect to the literature and our model. Section 3.5 presents the main results, and Section 3.6 proposes a discussion with additional measurements. Finally, Section 3.7 concludes the paper.

3.2 Related work

3.2.1 Explainable Artificial Intelligence

The “black-box” nature of ML algorithms (and other algorithms using AI) can make it difficult for end users, and sometimes even the producers themselves, to understand their processes and predictions (Castelvecchi, 2016). This opacity has led to a demand for transparency from stakeholders (Preece et al., 2018), regulators³ and end-users. To address these concerns, the field of XAI has emerged as a research area aiming to “make the results of AI systems more understandable to humans” (Adadi & Berrada, 2018).

The aim of XAI is manifold: it aims to make the decision-making process of AI models more transparent (Cheng et al., 2019), to help end-users and stakeholders understand how and why an AI has made a specific decision (Cai et al., 2019), and to build trust in the AI system (Shin, 2021; Weitz et al., 2021). To this end, several methodologies have been used in the research field. Some articles have focused on defining a framework for what a satisfactory explanation might look like from a human perspective. Riedl (2019) defines what might be a “human-centered” AI and explains that a key feature of XAI is to make it understandable when it fails or provides a result that “confuses” us. In addition, Weitz et al. (2021) explored the potential of virtual agents providing explanations of how an AI works for humans and gathered feedback from study participants. They found that end-users want in-depth explanations and to interact with the virtual agent itself, just as they would with a human being explaining something to them.

Another part of the literature focused on how these explanatory features could be implemented technically in relation to the AI model itself. Anjomshoe et al. (2019) proposed a framework describing three stages of explanation, namely “Explanation Generation”, “Explanation Communication”, and “Explanation Reception” (with several sub-stages for each stage). They classify—among other things—existing

³Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

XAI articles into several categories of technical features used to provide explanations of AI models. Another framework was proposed by Arrieta et al. (2020), offering an overview of existing AI models in the literature (such as neural networks or decision trees) and establishing a technical framework on how to provide model-dependent explanations. For example, they explain that some models can be explained ex-ante without any ex-post explanation of the results, while other models can only be explained ex-post (the interpretability of the model itself being too complicated). Finally, Beaudouin et al. (2020) propose another three-stage technical approach: contextual factor, technical solutions, and explainability parameters. This literature overall provides definitions, goals, and technical frameworks of the field of XAI. In the next sections, we will label this literature as “explainability”.

3.2.2 Explanations

Following Weitz et al. (2021), we distinguish, within the XAI literature, between the “explainability” literature—comprising the conceptual and technical framework presented earlier in Section 3.2.1—and the “explanations” literature, which evaluates the effects of AI model explanations on different outcomes and applies them to various fields. Indeed, since its emergence in the 1970s, the field of XAI has been applied to many other fields of research that use AI and/or ML in their processes (Beaudouin et al., 2020). This literature mainly comes from the fields of Computer Science (CS) and Human-Computer Interaction (HCI).

Some of this explanation literature focuses on the human-algorithm interaction. Cai et al. (2019) evaluate the effects of example-based explanations on users’ understanding of the results of an image recognition algorithm. Their results show that example-based explanations generally increase comprehension, but different types of examples can lead to varying effectiveness. Shin (2021) studies the effects of two sets of explanations of an algorithm on the understanding of non-expert stakeholders. The paper demonstrates that explanations overall improve objective and self-reported understanding of the algorithm’s output, but shows no difference in understanding between a “white-box” and “black-box” explanation model. Furthermore, although user understanding of the algorithm increased, their confidence in it did not follow the same trend. Weitz et al. (2021), meanwhile, propose a study measuring the effect of implementing virtual agents providing explanations of a sound recognition algorithm, on users’ confidence in that algorithm. They found that user confidence increases with the implementation of an explanatory virtual agent, but the magnitude of this effect depends on the number of features of such a virtual agent.

Some work has focused on social media and advertising, and how they relate to algorithmic explanations. Rader et al. (2018) applied an explanation methodology to social media by conducting an experiment using different ways to explain how Facebook’s News Feed algorithm works. They found that

each type of explanation increased users’ understanding of the algorithm and helped them determine the correctness and possible biases of such an algorithm. Eslami et al. (2018) exposed users to explanations of an advertising algorithm and found that this has increased their understanding of it and led to a form of “algorithmic disillusionment”: it updated the user’s beliefs about the algorithm’s performance. Finally, the explanation literature addresses the issue of algorithmic recommendations. Herlocker et al. (2000) provided experimental results demonstrating that explanations can improve acceptance of automated collaborative filtering algorithms (i.e., recommendations derived from the behavior of other users). Cramer et al. (2008) obtained similar results: explaining why a recommendation was made improves its acceptance by users. Overall, the literature on explanations can help demystify AI for non-technical users, leading to better understanding, acceptance, and adoption of AI technologies. These effects nevertheless vary from study to study, depending on the explanation methodology and the domain in which the explanation is carried out.

The results presented in this literature, in particular those demonstrating an increase in confidence in the results of algorithms, can be associated with the economic literature on the delegation of decisions. Indeed, this literature shows that an increase in user confidence in the decision-making process leads to an increase in the delegation of economic choices, whether theoretical (Bloom et al., 2012; Gur & Bjørnskov, 2017), or empirically measured onto several sectors (Guiso et al., 2008; Cingano & Pinotti, 2012; Özer et al., 2018; Holzmeister et al., 2023).

3.2.3 XAI, explanations, and music

The literature on XAI, whether on the explainability side (see Section 3.2.1) or on the explanations and applications side (see Section 3.2.2), has addressed the case of music streaming and its uses of AI, in particular the personalized recommendations via MRS. First, the definitions and aims of explainable recommendations (explainability) were presented as follows: “to make clear by giving a detailed description” (Tintarev & Masthoff, 2015) and “an explainable recommendation aims to answer the question of why” (Y. Zhang & Chen, 2020). This framework was completed by Afchar et al. (2022) and adapted specifically for MRSs by adding the notions of “fair recommendations” (i.e. addressing the problem of social biases in the recommendation) and “additional information”. In addition, they address the technical issues that can arise when explaining an MRS, including the choice of explanation methodology for conducting an experiment, and the engineering issues underlying real MRSs within the industry. There may be a trade-off between offering comprehensive explanations of all the features used by MRSs and the practical limitations faced by researchers. These limitations include data availability, replication costs for external researchers (those not affiliated with streaming services), and concerns about intellectual property for in-house researchers. In addition to this framework analysis for Explainable MRSs, Martijn et al. (2022)

integrate the need for cognition for different users in the design of an efficient MRS explanation.

Concerning the applications of explanations to MRSs, several studies have been conducted to measure the effects of such explanations on different outcomes. MRS explanations can help users better understand how recommendations were generated (Schröder & Ghajargar, 2021). Other work has proposed measuring outcomes in terms of behavior and human-computer interaction (Andjelkovic et al., 2016; Zhao et al., 2019), demonstrating overall that MRS explanations can lead to better user engagement. This article is in line with the literature on explanations. Although we do not deal with the question of the explicability of MRSs, we nevertheless adopt a framework already defined by this literature (see Section 3.4.2). Indeed, our aim is to measure the potential change in users' music consumption behavior (measured through various behavioral variables presented in Section 3.4) when MRS explanations are provided. The results provided could extend the literature on explanations applied to MRS, by documenting several human-computer interactions focusing on economic behaviors in the decision-making process of MRS users when they obtain additional information (see Section 3.3). Finally, this paper could provide insights for the music industry and the regulator, on discovery behaviors in the face of algorithmic recommendations, in particular, the evolution of these behaviors facing a less "black-box" nature of an MRS.

3.3 Theoretical framework

3.3.1 Setup

This Section aims to develop a tractable stylized model of music discovery behaviors in the presence of an MRS that guides our experimental strategy. The model focuses on beliefs and the effect of explanations. Importantly, this theoretical framework is tailored for music recommendation algorithms rather than all recommender systems. Music discovery often involves an iterative process in which users repeatedly interact with recommendations to refine their understanding of their own preferences and the system's patterns. Additionally, music discovery is inherently exploratory, as users may not have a fixed objective but instead seek serendipitous encounters with new artists or genres (chance discovery). Finally, music recommender systems often operate as black-boxes, creating information asymmetry between the system and users, which directly influences belief formation and the trade-offs modeled in our framework.

Agents We consider a simple framework, where future recommendations made by the MRS are not influenced by agents' listening behaviors. Therefore, recommendations are generated ex-ante and are not modified ex-post. We define a model of music consumption with two types of agents. Type 1 agents do not receive any explanations on the recommendation generated by the MRS, while type 2 agents receive information on the recommendation process. Regardless of the agent type, each agent

faces a trade-off between consuming the recommended track or exploring the next ones. This trade-off is inspired by Heidhues et al. (2021), who posits that when investigating available products, a consumer faces a trade-off between browsing and studying.⁴ We argue that due to the complexity and black-box nature of MRSs, agents face an agency problem due to information asymmetry and cannot verify whether the recommendations align with their preferences when evaluating the true relevance of recommended tracks. Therefore, we posit that they can dedicate time to listen to the recommended track to assess the quality of the recommendation, facing the possibility that the recommendation does not satisfy their preferences. Alternatively, they can dedicate an amount of time exploring the other recommended tracks but potentially allocate a significant amount of time until they find a recommendation that satisfies their preferences. We denote x as the time an agent spends consuming the recommended track and z as the time spent exploring other recommended tracks. Each agent is constrained by a total discovery listening time T , which they can allocate between x and z . Finally, we denote α and β , the weighting parameters assigned respectively to the consuming and exploring behaviors in the utility function of agents of types 1 and 2.

Music Recommender System We denote $I \geq 2$, a set of recommended tracks, and at the beginning of each period, each agent is assigned to a recommended track i with $i \in I$ determined based on their music consumption and habits.

Beliefs We suppose that our weighting parameters α and β may vary considering the beliefs on the MRS. We denote the initial beliefs of the agent of type 1 as θ , with $\theta \sim \mathcal{N}(\mu, \sigma^2)$. We consider providing agents of type 2 with explanations of the recommendation process of the MRS as an exogenous belief shock that results in updated beliefs denoted θ' , with $\theta' \sim \mathcal{N}(\mu', \sigma'^2)$. We suppose that although both beliefs are normally distributed across the population, the updated beliefs θ' may differ from the initial beliefs θ such that $\mu \neq \mu'$ and $\sigma^2 \neq \sigma'^2$.

Assumption 1 *Regardless of the agent's type, we suppose that the initial beliefs θ and the updated beliefs θ' impact the weighting parameters associated with the consuming and exploring behaviors α and β . Consequently, the utility functions for the two agent types are:*

⁴“For most of the paper, we posit that when investigating available products, a consumer faces a trade-off between browsing and studying. We have not found direct evidence of this trade-off. But given the self-evident observation that consumers must make decisions on how carefully to evaluate a complex product, the trade-off arises if consumers exhibit increasing marginal search costs.” Heidhues, P., Johnen, J., & Köszegi, B. (2021). Browsing versus studying: A pro-market case for regulation, *The Review of Economic Studies*, 88(2), 711.

1. For Agents of Type 1:

$$\begin{aligned}
 U_1(\theta) &= \alpha_1(\theta) \cdot x + \beta_1(\theta) \cdot z \\
 &\text{subject to } x + z \leq T \\
 EU_1 &= \sum_{i=1}^n U_1(\theta_k) \cdot \mathbb{P}(\theta_k)
 \end{aligned} \tag{3.1}$$

2. For Agents of Type 2:

$$\begin{aligned}
 U_2(\theta') &= \alpha_2(\theta') \cdot x + \beta_2(\theta') \cdot z \\
 &\text{subject to } x + z \leq T \\
 EU_2 &= \sum_{i=1}^n U_2(\theta'_k) \cdot \mathbb{P}(\theta'_k)
 \end{aligned} \tag{3.2}$$

where the expected utility for agents of type 1 quantifies the expected satisfaction from consuming and exploring, based on the beliefs θ , their variability k , and their associated probabilities $\mathbb{P}(\theta_k)$. The expected utility for agents of type 2 quantifies the expected satisfaction from consuming and exploring, based on the updated beliefs θ' , their variability k , and their associated probabilities $\mathbb{P}(\theta'_k)$.

3.3.2 Hypotheses

In this Section, we present the two hypotheses we formulate with our model that are tested in Section 3.5 using data from the experiment.

Hypothesis 1 *For type 1 agents, their propensity to explore, denoted by β_1 , is greater than their propensity to consume, denoted by α_1 .*

$$\alpha_1(\theta_k) < \beta_1(\theta_k) \tag{3.3}$$

On the one hand, Hypothesis 1 supposes that the information asymmetry caused by the black-box nature of the recommendation algorithm leads agents of type 1 to explore more over the recommended tracks, to gather information in order to update their initial beliefs θ on how the MRS behaves. In other words, prior beliefs are updated through learning recommendation patterns through the set of recommended tracks.

Hypothesis 2 *Conversely, for agents of type 2, the propensity to consume α_2 is greater than the propensity to explore β_2 .*

$$\alpha_2(\theta'_k) > \beta_2(\theta'_k) \tag{3.4}$$

On the other hand, Hypothesis 2 assumes that explanations of the MRS inner working lead to an exogenous shock on beliefs θ' . Consequently, the information asymmetry on the recommendation process

is reduced and Type 2 agents can spend more time consuming the tracks since the information-gathering process is no longer needed to comprehend the recommendation procedure.

3.4 Experimental design

The primary component of our experimental protocol is a listening decision task with a duration of 20 minutes, during which participants engage with recommended music tracks. At the onset of each listening, participants receive the headline characteristics (i.e. title, artist, cover, and length) of the recommended tracks via an embedded display of the Spotify interface integrated into our decision interface (see Figure 3.A.2). The ultimate number of decisions made by participants is contingent upon their listening behaviors (within a limit of 50 recommended tracks).

3.4.1 Behavioral Predictions

Within the decision interface, participants have the option to listen to a recommended track for a self-determined duration or to proceed to the next recommended track. This mechanism serves to measure the outcomes of consuming/exploring behaviors, aligning with our theoretical model (see Section 3.3). Additionally, participants are asked to indicate whether they are already familiar with the recommended tracks, thereby controlling for the knowledge of the tracks and observing the actual discovery behaviors (“known” button). Furthermore, an implemented feature allows participants to express their preferences through designated “like” and “dislike” buttons, providing a means to signify their inclination or not towards tracks they find (dis)appealing. Importantly, the “known” and “like/dislike” decisions are not mutually exclusive, enabling participants to signal both their familiarity and appreciation for specific tracks. Decision outcomes and the corresponding predicted behaviors in accordance with our theoretical model are depicted in Table 3.1.

Table 3.1: Behavioral Predictions Table

		Listening Time (t_{ik})	
		$t_{ik} < t_0$	$t_{ik} \geq t_0$
Skip	Yes	z (i.e., exploring)	x (i.e., consuming)
	No	Does not apply	x (i.e., consuming)
Known $\times$ Skip	Yes	Excluded	Excluded
	No	Does not apply	x (i.e., consuming)

Note: With t_0 denoting the threshold at which a track is classified as “consumed”, based on the industry standard and generated data. Instances deemed “Does not apply” refer to cases that are logically impossible, such as when a participant does not skip a track but the relative listening time t_{ik} is still below t_0 . “Excluded” cases are not incorporated into the predictions, because “Known” tracks are not considered as discovered tracks.

3.4.2 Experimental treatments

Participants within the subject pool were randomly assigned to various treatments outlined in Table 3.2. The music preferences questionnaire (see Section 3.4.3) was administered across all groups, with specific details provided in the subsequent sections. Treatment allocations involved the utilization of a simplified MRS algorithm that generates a set of 50 recommended tracks (see Section 3.4.2). Recommendations were either randomly assigned based on the preferred music genre and the declared variety of tastes (*Algorithm*, *Explanations*), or conversely to these preferences (*Anti-Algorithm*).

Within the theoretical framework, the *Algorithm* and *Explanations* conditions follow naturally from the model’s assumptions about belief formation and the role of information asymmetry. In the *Algorithm* condition, recommendations align with participants’ stated preferences, meaning agents engage with tracks that are likely to match their tastes. This setup reflects a baseline scenario in which users interact with a music recommender system (MRS) under standard conditions—without additional information about how the recommendations were generated. The *Explanations* condition acts as the exogenous belief shock, as participants receive information on the recommendation process, reducing information asymmetry. According to the model, this shock should influence agents’ weighting parameters for consumption and exploration, leading them to rely more on the recommendations provided.

In contrast, the *Anti-Algorithm* condition is implemented to represent the highest possible misalignment between user preferences and algorithmic recommendations. Specifically, in this treatment, recommended songs explicitly conflict with participants’ stated music preferences. This approach intentionally simulates scenarios in which recommendation systems significantly diverge from consumer preferences due to either poor personalization, systematic algorithmic biases, or misaligned commercial objectives, such as the promotion of specific artists or genres. While one might suggest that randomly assigning songs could serve as an alternative control condition, we argue against this approach on theoretical and practical grounds. Randomized song recommendations do not replicate genuine human-algorithm interactions, as they lack any algorithmic logic or personalized recommendation process. Instead, the *Anti-Algorithm* condition deliberately introduces systematic but controlled variation, which enables clearer interpretation of user reactions to preference misalignment. Consequently, despite its counterintuitive nature, the *Anti-Algorithm* treatment provides a conceptually richer baseline for examining the behavioral impact of algorithmic recommendations that fundamentally deviate from users’ preferences. Additionally, this approach allows us to control for the potential influence of the algorithm’s quality for the *Algorithm* and *Explanation* conditions.

Conventional music recommender systems incorporate diverse inputs to formulate personalized recommendations, encompassing factors such as music genres (including audio features), artists, release dates,

Table 3.2: Conditions Overview

Condition	Recommendations	Information on the recommendations
<i>Anti-Algorithm</i>	Conversely to the stated preferences	None
<i>Algorithm</i>	Based on the stated preferences	None
<i>Explanations</i>	Based on the stated preferences	Receive Information

etc. Moreover, many MRSs integrate proprietary metrics developed by private streaming companies, such as listening mood or social listening context. Advanced MRSs employ Machine Learning (ML) techniques, such as Neural Networks, to generate predictions—specifically, recommended tracks—based on individual and aggregated listening history. These sophisticated systems can update recommendations dynamically in response to recent listening activity (Afchar et al., 2022). While it could be argued that explaining the inner workings of sophisticated MRSs might prompt consumers to manipulate their listening history for improved recommendations, this study focuses on assessing the *ex-post* impact of explaining MRS recommendation processes on music discovery behavior rather than its *ex-ante* impact, which could involve hypothetical listening history manipulation. Also, we postulate that manipulating such algorithms would necessitate sustained efforts over an extended period, rather than being achievable in a single listening session of 20 minutes.

This study acknowledges the sensitive nature of privately held metrics and listening data, as well as the intricate workings of ML models, shaping the decision to opt for simplified algorithms aligned with industry standards. Our algorithm has generated recommendations based on the music genre criterion (see Section 3.4.2). Recommendations were finalized at the end of the music preferences questionnaires, relying on elicited preferences and consumption habits rather than individual listening histories.

Explanation methodology for Music Recommender System

As previously discussed in Section 3.2.3, Afchar et al. (2022) offer a comprehensive review of explanation methods for MRSs. Their work primarily inspires the explanation methodology used for this experiment. Before delving into explanation methods for MRSs, Afchar et al. (2022) provide an overview of existing explanation methods for recommender systems (RSs), inspired by the six RS explanation types outlined by Y. Zhang and Chen (2020) as presented in Table 3.3. The algorithm developed for this experiment relies on music features and diversity of tastes (or multimodal information Afchar et al. (2022)) to generate personalized recommendations.

Given this context, our explanation methodology draws on existing literature on feature-based explanations for MRSs. Afchar et al. (2022) assert that feature-based explanations are meaningful only if the features themselves are interpretable, which is highly the case for our situation. However, current

Table 3.3: RSs Explanation Types

Methods	Principles of the explanation
<i>Example-based</i> explanations	Similarity of the item with other previously liked items by the user
<i>Feature-based</i> explanations	Item features presented as tags relevant for the user
<i>Opinion-based</i> explanations	Relevant aspects of recommended items enriched with sentiment based on reviews
<i>Sentence</i> explanations	Predefined sentences with placeholders for features eventually tailored for the user or based on the recommendation
<i>Visual</i> explanations	Visual elements (image regions or caption words)
<i>Social</i> explanations	User’s friends choices regarding recommended items

limitations, such as the assumption of a limited number of interacting features (J. Chen et al., 2018) or feature independence (Ribeiro et al., 2016a), can restrict the efficiency of feature-based explanations. When designing our algorithm, we assumed a limited correlation between our two main features. This assumption aligns with the simplicity of our recommender system, minimizing potential strong interactions between selected features and making feature-based explanations appropriate.⁵

Besides using features in our explanations, we also disclosed the algorithm’s inner workings to participants by introducing them to its primary optimization parameters—audio features⁶—used to generate recommendations. While audio features could be considered “sub-features” strongly associated with musical genres, they were crucial in distinguishing how the algorithm recommended tracks. Therefore, to provide the most comprehensible explanation, we also embedded the algorithm’s inner workings in our explanation strategy. Lastly, we provided examples of the algorithm’s recommendation process for specific cases and a graphical explanation personalized based on questionnaire responses used by the algorithm (Figures 3.1 and 3.2).

The choice of a global explanation over local ones is favored due to its ease of implementation and its focus on explaining the inner workings of the algorithm and the main features used, while also avoiding redundancy. In contrast to local explanations, which would require generating item-specific justifications for each recommendation, a global explanation provides participants with an overarching understanding of how the system operates, what information it relies on, and how these factors shape the set of recommendations they encounter. This approach not only reduces the cognitive load for participants but also ensures that they grasp the main logic of the recommender system before interacting with it. Importantly, this strategy mirrors practices in the industry, where global explanations are often preferred over local ones because they are more scalable and easier to communicate consistently to a large and diverse audience.

⁵As confirmed by a χ^2 test, there is indeed a significant association between our two features (p-value < 0.05). However, the effect size measured by Cramér’s V is approximately 0.258, indicating a weak to moderate association. Detailed results and additional analysis are provided in Appendix 3.A.1.

⁶“Audio features” are tracks’ sonority metrics generated by Spotify’s audio engine. For additional details on these features, see Spotify for developers.

As illustrated by Spotify’s own communication (Figure 3.A.1), industry leaders typically disclose the broad principles driving their recommendation engines rather than providing item-by-item explanations. Spotify emphasizes that recommendations are based on a combination of personal listening history, contextual factors (time of day, device, user actions), social information (follows, trends, similar listeners), content-level metadata (genres, release dates), and even commercial considerations (such as promotional tools for artists and labels). This type of global explanation offers users a transparent yet high-level view of the multiple signals involved, without overwhelming them with details for each specific recommendation. Our methodological choice to adopt a global explanation thus aligns with these industry practices.

Figure 3.1: Wording of the explanations presented to the treated group

Explanation of how the algorithm works

The algorithm uses your **declared favorite musical genre** and your **stated variety of tastes*** to generate a recommendation.

The recommendation is **based on your declared favorite musical genre**, Hip-Hop or Rap, as well as **sounds that are close to or far** from this genre, depending on the variety of your tastes.

Sonorities are characteristics specific to each track and can be quantified. They include, for example, instrumentality (the combination and use of instruments and voices to create the track) and valence (the degree of positive or negative emotion evoked by the track).

***Variety modalities being the following: exclusive, not very varied, moderately varied, varied, and very varied.**

Examples:

(1) If you declare:

→ **That you prefer Jazz**

→ **That your tastes are not very varied**

Then the algorithm recommends you **jazz tracks or tracks with very similar sonorities**.

(2) If you declare:

→ **That you prefer classical music**

→ **That your tastes are very varied**

Then the algorithm recommends you **classical tracks and tracks with very different sonorities from this genre (like Rap or Pop music)**.

“Music Genre” Algorithm

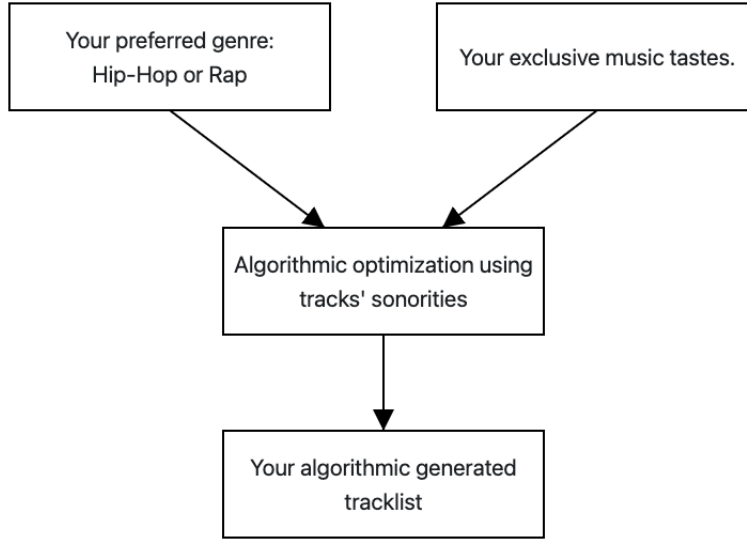
Our “Music Genre” algorithm⁷ employs two input variables to generate recommendations: the participant’s elicited preferred musical genre and the range of musical genres encompassed in their listening habits (variety of music tastes).

Let I denote the set of recommended tracks, with $i = \{1, \dots, n\}$, $i \in I$. We define :

$$I = \text{Pref}_k + \text{Other}_k \quad (3.5)$$

⁷For details on the “Music Genre” algorithmic recommendation process, see Algorithm 1.

Figure 3.2: Explanation diagram presented to the treated group



where Pref_k is the set of recommended tracks, with $\text{Pref}_k = \{p_{1k}, p_{2k}, \dots, p_{nk}\}$, randomly drawn from a dataset containing tracks belonging to the same genre as the participant's k preferred genre, and Other_k is the set of recommended tracks, with $\text{Other}_k = \{o_{1k}, o_{2k}, \dots, o_{nk}\}$, randomly selected from a dataset comprising tracks of other genres that have been selected by participant k according to the specific following rule.

For all instances o in Other_k , the recommended tracks are selected according to a rule that aims to minimize the number of recommended tracks and the distance between recommended tracks, based on participants' stated preferences:

$$\underset{\lambda}{\operatorname{argmin}}\{|\text{Other}| \leq \delta_k \cdot Y\} \quad (3.6)$$

such that $\forall o \in \text{Other}, \forall j$:

$$\mu_{jk} - \lambda \cdot \sigma_j \leq f_{jn} \leq \mu_{jk} + \lambda \cdot \sigma_j \quad (3.7)$$

where f_{jn} represents an audio feature j of a track n , μ_{jk} denotes the average value of the audio feature j in the participant's k preferred genre and σ_j represents the standard deviation of these average values of audio features j across all music genres. λ is the parameter that minimizes the length of Other , and δ is the diversity propensity coefficient determined by the participant's k declared variety of tastes that determines the length of Other . Finally, Y corresponds to the minimal size of Other (the minimum possible length of Other being 25). This algorithm returns 50 tracks, 25 randomly drawn from Pref_k and 25 randomly drawn from Other_k .

“Anti-Algorithm”

Building on previous examples of experimental designs utilizing “reverse” treatments (Cooper et al., 2018), we develop a “Anti-Algorithm” alongside the “Music Genre” algorithm⁸ to recommend tracks to the *Anti-Algorithm* group. The “Anti-Algorithm” is designed to generate recommendations by reversing the participant’s stated music preferences. This means that for each participant, tracks are partially selected from the genre that is the most distant from their stated preferred genre. Additionally, using the range of musical genres encompassed in their listening habits, we recommend tracks that are opposite to the stated variety of music tastes.

Let I denote the set of recommended tracks, with $i = \{1, \dots, n\}$, $i \in I$. We define :

$$I = \text{RevPref}_k + \text{Other}_k \quad (3.8)$$

where RevPref_k is the set of recommended tracks, with $\text{RevPref}_k = \{p_{1k}, p_{2k}, \dots, p_{nk}\}$, randomly drawn from a dataset containing tracks belonging to the furthest away genre as the participant’s k preferred genre.

We use audio features to determine the furthest away genre. First, we denote p_k , the participant preferred genre, for each genre m where $m \neq p_k$:

$$\Delta_{p_k, m, j} = |\mu_{p_k, j} - \mu_{m, j}| \quad (3.9)$$

with $\Delta_{p_k, m, j}$ the absolute difference between the average value of audio feature j of tracks in the participant’s preferred genre noted $\mu_{p_k, j}$ and the average value of audio feature j for the genre m ($m \neq p_k$) noted $\mu_{m, j}$.

We then sum these differences across all audio features j :

$$D_{p_k, m} = \sum_j \Delta_{p_k, m, j} \quad (3.10)$$

Therefore, the genre furthest away from the preferred genre of participant k , noted $\bar{p}_k$ is given by:

$$\bar{p}_k = \underset{m}{\operatorname{argmax}} D_{p_k, m} \quad (3.11)$$

$\text{Other}_k = \{o_{1k}, o_{2k}, \dots, o_{nk}\}$, randomly selected from a dataset comprising tracks of other genres that have been selected by participant k according to the specific following rule.

For all instances of o in Other_k follow the same rule presented in Section 3.4.2 with the exception

⁸For details on the “Anti-Algorithm” recommendation process, see Algorithm 2.

that it uses $\bar{\delta}$, the reversed diversity propensity coefficient elicited by the participant's k declared variety of tastes, to determine the length of Other:

$$\underset{\lambda}{\operatorname{argmin}}\{|\text{Other}| \leq \bar{\delta}_k \cdot Y\} \quad (3.12)$$

such that $\forall o \in \text{Other}, \forall j$:

$$\mu_{jk} - \lambda \cdot \sigma_j \leq f_{jn} \leq \mu_{jk} + \lambda \cdot \sigma_j \quad (3.13)$$

Employing the same parameters as the “Music Genre” algorithm, the “Anti-Algorithm” distinguishes itself by utilizing the variable $\bar{\delta}$ to generate inverse recommendations. Specifically, when participants report predominantly listening to a single genre, the algorithm is more likely to recommend tracks that deviate significantly from their stated genre preference. In contrast, when participants indicate diverse listening habits across multiple genres, the algorithm tends to recommend tracks that align more closely with their stated preferences. Finally, Y corresponds to the minimal size of Other (the minimum possible length of Other being 25). This algorithm returns 50 tracks, 25 randomly drawn from RevPref_k and 25 randomly drawn from Other_k .

3.4.3 Questionnaire on music preferences and consumption habits

The objective of the questionnaire on music preferences and consumption habits is two-fold. Firstly, the elicited preferences and habits serve as input variables for our algorithms (see Section 3.4.2), to generate a set of personalized recommendations. Consequently, a substantial portion of the questionnaire is tailored explicitly to complement the algorithmic framework, diverging from conventional practices in scientific literature or the music industry. Secondly, the data collected serves a pivotal role in our regression analysis, functioning both as control variables and explanatory variables. This second category of questions aligns with established literature on music preference elicitation and common practices within the cultural domain. The majority of our questionnaire is inspired by both the music and the sociodemographic sections of the latest survey on Cultural Practices (“*Enquête sur les Pratiques Culturelles*”) by the French Ministry of Culture in 2018. This survey went through 6 editions between 1973 and 2018.⁹ Some questions have been added to fit the experiment, mostly to elicit the attitude towards music discovery and consumption, others have been modified according to our music genre categorization. This questionnaire was divided into two parts: a part that the participants completed before the execution of the main experimental task (Table 3.A.2), and a second part that the participants completed after this task (Table 3.A.3).

⁹For any additional information on the survey, refer to: <https://www.culture.gouv.fr/espace-documentation/statistiques-ministerielles-de-la-culture2/L-enquete-pratiques-culturelles>.

The prior-to-task questionnaire elicits the participants’ music tastes to calibrate our algorithm and controls several characteristics. Two questions explicitly serve as input variables for our algorithm, namely the preferred genre and the variety of tastes. The other questions in the music preferences questionnaire (favorite artists, frequency of music listening, etc.) serve both to make it difficult for participants to identify directly which variables were being used to propose a personalized recommendation and as controls in our empirical analysis. To infer the differential impact of providing explanations within an MRS, we identified two key variables distinguishing participants regarding the “expert” status (in addition to the traditional socio-demographic variables such as age, gender, educational background, social and occupational category). Indeed, we formulate the hypothesis that both the use of and comfort with digital tools such as streaming services, and musical knowledge and practices may have an impact on how the explanations are received, therefore affecting the discovery behaviors. IT and computer knowledge questions sourced from EuroPass.¹⁰ The music knowledge questions mostly come from the Cultural Practices survey and cover topics such as instrument practice, listening devices, and cultural events. In addition to these questions, we added a discovery device question, asking the participants on which device they usually discover new tracks. Finally, two questions to elicit their mood in the questionnaire: one on the enjoyment of discovery in general, and the other on participants’ opinions of music streaming services. The purpose of these questions is twofold: firstly, to identify pre-existing attitudes towards streaming and discovery, and secondly, to control for any differences in these attitudes in our empirical strategy.

The post-task questionnaire consists of two sections. The first Section focuses on feedback about the main task, gathering participants’ general impressions of the recommended tracks and their understanding of the recommendation process. This helps identify participants who may not have fully understood the algorithm’s explanation and allows for an estimation of their beliefs about how the MRS functions. The second Section includes psychometric items to assess participants’ behaviors, intentions, and attitudes toward MRSs.

3.4.4 Tracks dataset

To proceed with the main task of the experiment (i.e., getting subjects to listen to and interact with music tracks), we need to rely on a dataset containing enough tracks to generate a set of 50 recommended tracks when the algorithms’ filters are applied (see Section 3.4.2). We therefore collected this data using Spotify’s API. We selected various categories on Spotify’s interface¹¹ according to the music genres proposed in the questionnaire (see Section 3.4.3). Next, we collected all editorial playlists (i.e. playlists created by Spotify’s editorial teams) for each category, resulting in a set of 654 distinct playlists. We

¹⁰See <https://europa.eu/europass/fr>.

¹¹See <https://open.spotify.com/search>.

then randomly selected 10 playlists for each genre to ensure that all musical genres had equal weight in the final dataset.

After removing missing values (mostly resulting from errors of API calls), the final tracks dataset contains 6,820 songs (of which 6,049 are unique) from 134 playlists. The main information we use in this dataset is the track IDs. We used these IDs to generate the display on our decision interface by making a request to Spotify’s embed functionality.¹² In addition to track IDs, we used the audio features of the tracks to calibrate our recommendation algorithms (see Section 3.4.2). We also collected the tracks’ metadata such as duration, track popularity, artist popularity, etc. These variables will serve as controls in our econometric analysis. Finally, we assigned a music genre to each track by creating a set of dummy variables corresponding to the 14 musical genres (Pop, Hip-Hop, Jazz, etc.) from which the tracks are coming. We associated these genres with regular expressions within the “artist genre” array and assigned the value 1 to the corresponding music genre variable (a track can be assigned to several genres). If no match is found, or if the “artist genre” array is empty, the track is assigned to the most common genre dummy variable in the playlist from which it comes.

Figure 3.3: Number of tracks per music genre in the tracks dataset

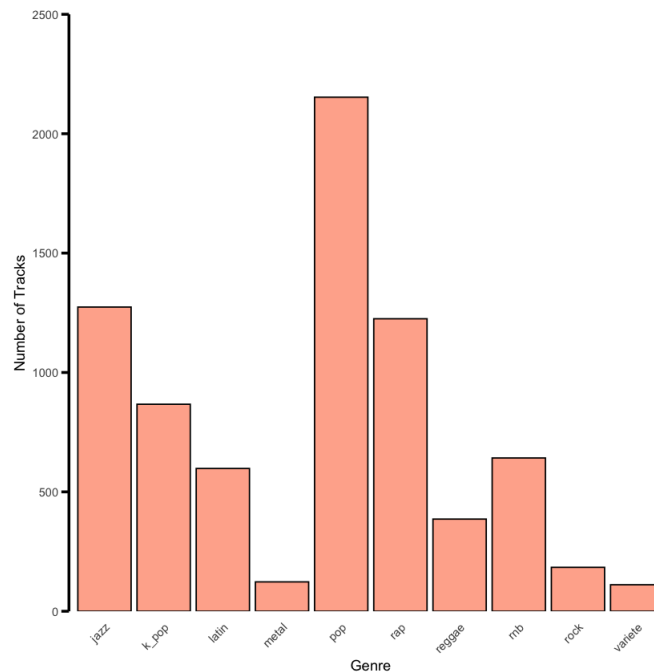


Figure 3.3 presents the number of tracks in our dataset per music genre. As tracks can be categorized into several genres, we observe a higher proportion of tracks from general genres or with multiple sub-genres, such as Pop, Hip-hop, or Jazz. Playlist length can also play a role in the distribution of tracks by genre. Table 3.4 presents the summary statistics of the tracks dataset, we observe a normal distribution

¹²See Spotify for developers.

of tracks and artist popularity.¹³ On the other hand, we observe a positively skewed distribution of the number of tracks by playlists, spanning from 1 track to 580, with a mean value of 50.89 and a median of 39.50.

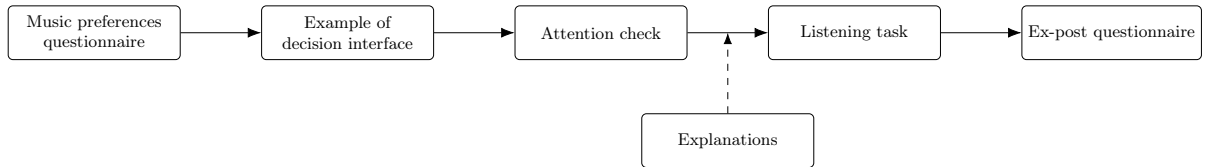
Table 3.4: Summary statistics of the tracks dataset

	Min	Max	Mean	Median	s.d.	Obs.
Track popularity	0.00	99.00	49.88	50.00	18.23	6,049
Artist popularity	0.00	100.00	44.55	44.00	17.96	3,988
Tracks duration (<i>in sec</i>)	17.50	984.65	194.64	187.00	57.98	6,049
Number of tracks per playlist	1.00	580.00	50.89	39.50	72.72	134

3.4.5 Laboratory Implementation

Personalization of recommended tracks is achieved through an algorithm that generates suggestions based on responses to a music preferences questionnaire administered prior to task commencement (see Sections 3.4.2 and 3.4.3). Upon eliciting participants' music preferences, task instructions were presented, followed by an illustrative example of the decision interface. Participants were only allowed to progress once they had successfully answered comprehension questions about the instructions. After the 20-minute task duration (or when the 50-track limit was reached), participants concluded the experiment and received a fixed payoff of 15 euros. They also had the option to ask for the "liked" tracks to be sent during the experiment. This methodology ensures a systematic and controlled approach to understanding participants' discovery processes in the context of personalized music recommendations and the effect of explaining the inner workings of an MRS (Figure 3.4).

Figure 3.4: Flowchart of the experiment



The laboratory experiment was conducted at the Laboratoire d'Economie Expérimentale de Paris,¹⁴ the experimental economics facility affiliated with the University Paris 1 Panthéon-Sorbonne and the Paris School of Economics. Participants were recruited from a database of volunteer participants. Additionally, to avoid any problem in the listening interaction, an exclusion criterion was applied to individuals

¹³This normal distribution is also reflected in a Kolmogorov-Smirnov test for both variables that indicates a normal distribution with $p < 0.001$.

¹⁴<https://leep.univ-paris1.fr>

with hearing difficulties. The experimental sessions took place between June 6th and July 5th, 2024. Instructions¹⁵ were delivered to participants at the beginning of each session. Subsequently, participants accessed the local software, developed in oTree (D. L. Chen et al., 2016). Following the completion of the questionnaire, participants were given (or not) the MRS explanations depending on the treatment (see Section 3.4.2), and started the listening decision task. In total, 364 participants completed the experiment, spread over 22 sessions, it was conducted in French and lasted an average of 40 minutes. Participants received a fixed payment of €15.00, equivalent to approximately US\$16.41 based on the exchange rate in July 2024. Treatment allocation was randomized and made as follows: 133 participants in the *Explanations* condition (36.54%), 113 in the *Algorithm* condition (31.04%), and 118 in the *Anti-Algorithm* condition (32.42%). The order of the sessions was randomly determined, and consequently, the distribution of the number of sessions across experimental conditions was also randomized. We performed a balance test to ensure the randomization across groups was effective (Table 3.A.7). The sample size was determined ex ante based on budgetary constraints, which set the maximum number of participants that could be recruited.

3.5 Results

3.5.1 Data and Empirical strategy

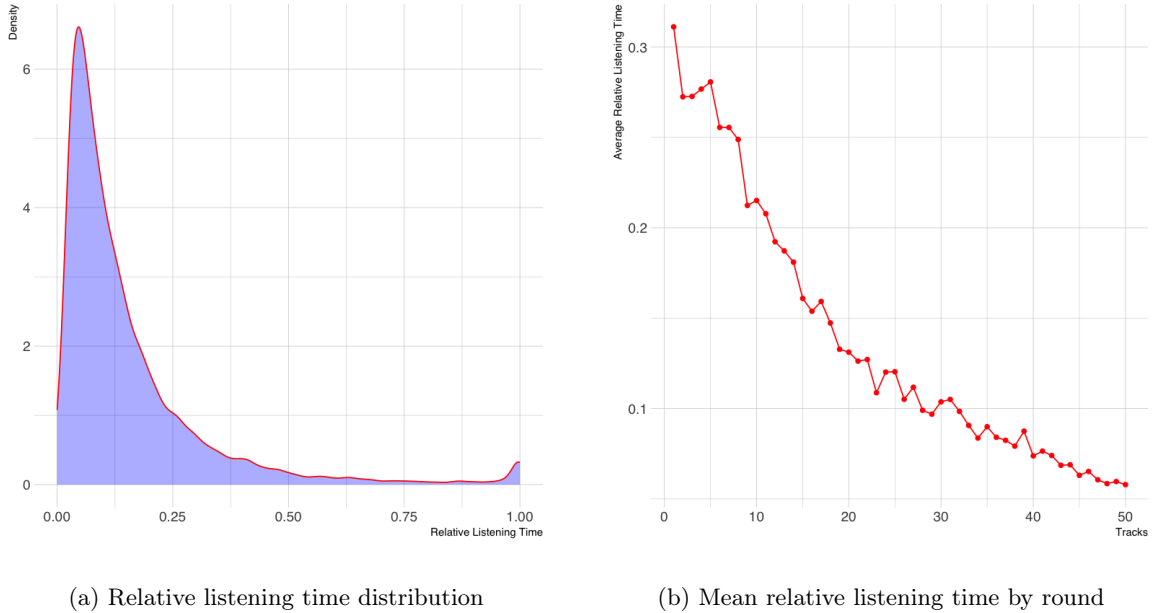
Table 3.5: Summary statistics of control variables and dependent variable

	Min	Max	Mean	Median	s.d.	Obs.
Number of listened tracks (n_k)	6.00	50.00	37.59	45.00	14.44	364
<i>Anti-Algorithm</i>	6.00	50.00	37.04	44.00	14.45	118
<i>Algorithm</i>	6.00	50.00	36.37	44.00	15.17	113
<i>Explanations</i>	6.00	50.00	38.84	48.00	13.74	133
Number of known tracks (k_k)	0.00	45.00	4.77	3.00	6.20	364
<i>Anti-Algorithm</i>	0.00	28.00	3.53	2.00	5.12	118
<i>Algorithm</i>	0.00	24.00	4.86	3.00	5.71	113
<i>Explanations</i>	0.00	45.00	5.80	3.00	7.24	133
Relative listening time (t_{ik})	0.00	1.00	0.15	0.10	0.17	11,946
<i>Anti-Algorithm</i>	0.00	1.00	0.16	0.10	0.17	4382
<i>Algorithm</i>	0.00	1.00	0.15	0.09	0.17	4126
<i>Explanations</i>	0.00	1.00	0.14	0.09	0.15	5177
Absolute listening time	0.00	390.00	27.96	18.00	31.08	11,946
<i>Anti-Algorithm</i>	0.00	284.00	28.17	19.00	30.89	4382
<i>Algorithm</i>	0.00	308.00	29.05	19.00	33.39	4126
<i>Explanations</i>	0.00	390.00	26.90	18.00	29.22	5177

¹⁵The instructions are available in Table 3.A.4.

On average, study participants were 41 years old ($SD = 17$ years), with a majority of women, representing 64.01% of the sample. Level of education varied: 31.32% held a Master’s degree, 39.28% a Bachelor’s degree or University Technology Diploma (DUT), and 25.55% a high school diploma. Students represent 26.92% in the dataset, employees 24.17%, executives and highly skilled workers 14.56%, intermediate professions 8.52%, other professions 7.42%, and craftsmen, shopkeepers, and company directors 4.12%.¹⁶ As explained in Section 3.4.3, we have identified two key variables likely to influence the effect of the MRS explanation: “expert” status in numerical knowledge and musical knowledge. To construct these variables, a set of questions is asked on these two subjects. The main components of differentiation between participants are determined by PCA¹⁷ and then summed to create a “knowledge score” for each participant, in both numerical and musical knowledge. The aim of this article is to examine the influence of explanations on music discovery behavior. Thus, we excluded the recommended tracks that participants had already reported as known (via the “known” button, see Section 3.4). On average, participants listened to 37.59 tracks during the listening decision, and knew 4.77 of them, giving a knowledge rate of 12.68%. In total, participants listened to 13,685 tracks, of which 1,739 were known, leading to a final number of observations in our dataset of 11,946 tracks. Descriptive statistics for these variables are presented in Table 3.5.

Figure 3.5: Relative listening time distribution and evolution across rounds



Our main dependent variable is built as follows: for every recommended track, the time spent dedicated to listening to it is recorded, and divided by the total length of the track (retrieved from the tracks dataset

¹⁶ Additional results on numeric and character variables are available in Tables 3.A.5 and 3.A.6.

¹⁷ To reduce the dimensionality, we selected the lowest number of components that capture 80% of the variance.

and labeled as “duration_ms”, see Section 3.4.4). Thus, we create a measure of relative listening time for every track¹⁸ we denote as t_{ik} , in alignment with our theoretical framework and experiment design (see Section 3.4.1). Descriptive statistics for this variable are available in Table 3.5: we observe a positively skewed distribution, confirmed by the density plot (Figure 3.5a). We also observe the average relative listening time across all participants dropping as the rounds progress (Figure 3.5b), in line with our expectations and consistent with the literature, which suggests that repeated exposure to similar tasks or stimuli often leads to fatigue or diminishing engagement over time (Kahneman, 1973; Hidi & Renninger, 2006; Charness & Gneezy, 2009; DellaVigna, 2009). In the subsequent analyses, the results will be controlled by the round to account for this diminishing engagement, and therefore, observe a consistent effect of MRS explanation across all observations.

Our empirical strategy relies mainly on the following linear model:

$$t_{ik} = \beta_0 + \beta_1 \text{Algorithm}_k + \beta_2 (\text{Algorithm}_k \times \text{Explanations}_k) + \beta_3 X_k + \beta_4 N_i + \beta_5 V_{ik} + \epsilon_{ik} \quad (3.14)$$

where t_{ik} is the relative listening time of a track i for a participant k , *Algorithm* is a dummy variable corresponding to the algorithm type that generated the recommendations of participant k (1 if the “Music genre” algorithm is used, 0 if the “Anti-Algorithm” is, see Section 3.4.2), *Explanations* is a dummy variable corresponding to the treatment of a participant k (1 if k is treated, 0 if not). X is a vector containing a set of participant k control variables (age, gender, education, opinion on streaming services, digital and musical knowledge), N is a vector controlling for the track i and V is a vector controlling for the track’s ik characteristics.¹⁹ The rationale behind the use of the interaction term *Algorithm* $\times$ *Explanations* is to allow a comparison between the *Explanations* and the two control conditions: the parameter β_1 estimates the difference between the *Anti-Algorithm* and the *Algorithm* conditions, and the parameter β_2 estimates the incremental effect of the treatment between *Algorithm* and the *Explanations* conditions. When combining β_1 to β_2 , we can estimate the difference between the *Explanations* and the *Anti-Algorithm* conditions.

3.5.2 Treatment effect

Descriptive results reveal a difference in relative listening between experimental conditions: MRS explanations appear to have a negative impact on t_{ik} , as illustrated in Figure 3.6. Mean relative listening time varies significantly between groups, with values of 0.156 for the *Algorithm* condition, 0.164 for the

¹⁸The relative time measure is capped at 1.00 for ease of interpretation, this value may be greater than 1 if the participant has gone backward while listening to the track, bringing the total listening time to a duration greater than the track itself.

¹⁹The tracks characteristics are the variable outlined in Section 3.4.4, namely tracks’ popularity, artists’ popularity, tracks’ duration and audio features.

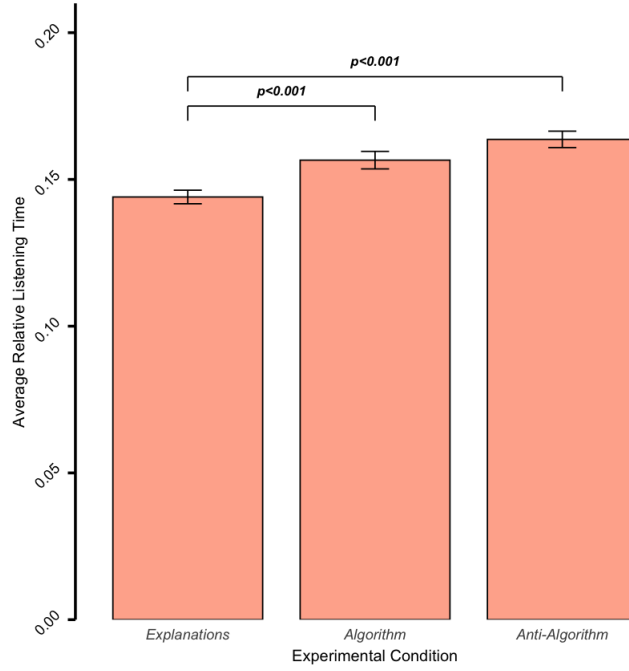


Figure 3.6: Average relative listening time (t_{ik}) by experimental condition

Anti-Algorithm condition, and 0.144 for the *Explanations* condition. This first finding contradicts the predictions of our model, suggesting that MRS explanations tend to reinforce the “exploring” rather than the “consuming” behavior, leading to a decrease in average t_{ik} . Moreover, the Anti-algorithm condition displaying higher listening time may seem contradictory with our initial hypotheses, but appears consistent with the previous literature on music recommendation (Adomavicius et al., 2018), which shows that recommendations (even biased) increase the willingness to pay for the tracks (as a proxy for consumption).

When controlling for the variables outlined in equation 3.14, the effect of the MRS explanation on t_{ik} becomes statistically insignificant. As shown in Table 3.6, the average relative listening time is 0.016 (at the 10% confidence level) lower for participants who received explanations (*Explanations* group), compared to the two control groups (*Algorithm* and *Anti-Algorithm*). This lower relative listening time remains negative and significant when participants, round, and track controls are included (*Model 2*). When adding the algorithm control, the treatment effect disappears: we observe no statistical difference in relative listening between *Algorithm* and *Explanations* groups.²⁰ The negative differences in Models 1 and 2 are driven by the higher relative listening of the *Anti-Algorithm* group, rather than a detrimental impact of explanations. This outcome contradicts our initial hypothesis, which posited that explanations could impact the allocation of time between “consuming” and “exploring”, thereby influencing t_{ik} .

²⁰To benchmark these null results, we compute ex-post MDEs with clustered SEs: at 80% power and $\alpha = 0.05$, we could detect ≥ 0.026 for the *Algorithm* main effect and ≥ 0.023 for *Algorithm* $\times$ *Explanations* (Table 3.6). The corresponding estimates (-0.006 ; -0.010) lie below these thresholds, and their 95% CIs $[-0.025, 0.012]$ and $[-0.026, 0.006]$ exclude only larger effects. Restricting to *Algorithm* vs. *Explanations*, we detect no effect of explanations on relative listening; given the MDEs, we can rule out effects larger than ~ 0.02 - 0.03 units (0.11-0.19 SD) but not smaller ones.

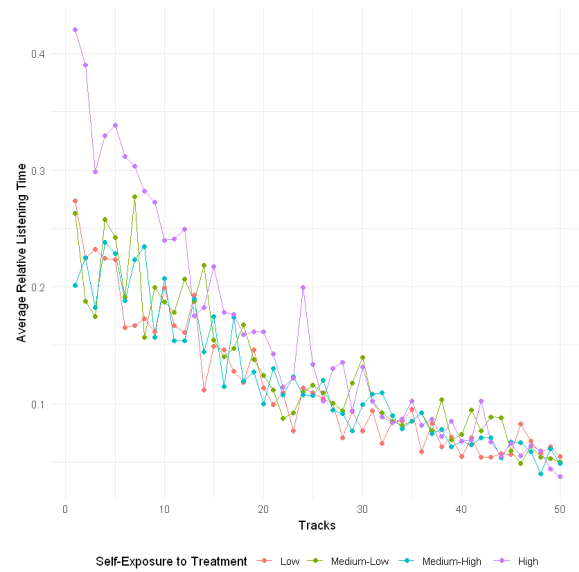
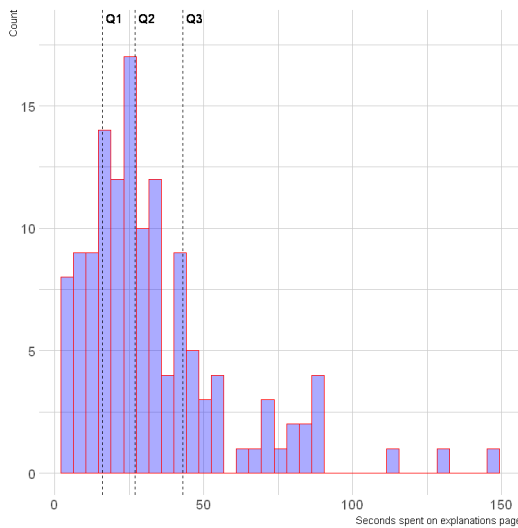
Table 3.6: Explanations' effect on Relative Listening

	<i>Dependent variable:</i>		
	(1)	Relative listening (t_{ik}) (2)	(3)
Control groups	Ref.	Ref.	
Explanations	-0.016* (0.009)	-0.013** (0.007)	
Anti-Algorithm			Ref.
Algorithm			-0.006 (0.009)
Algorithm $\times$ Explanations			-0.010 (0.008)
Algorithm control	No	No	Yes
Participants controls	No	Yes	Yes
Round FE	No	Yes	Yes
Observations	11,946	11,946	11,946
R ²	0.002	0.238	0.238

Note: Clustered standard error at the participant level in parentheses. Participants' controls include: gender, age, digital PCA, music PCA, discovery mood, and streaming opinion. Track controls include: artist's popularity, track's popularity, track duration, and audio features.

*p<0.1; **p<0.05; ***p<0.01

3.5.3 Self-exposure to treatment



(a) Distribution of the Time spent on the Explanations page

(b) Mean relative listening time by round and by self-exposure to treatment

Figure 3.7: Self-exposure to treatment distribution and evolution across rounds by group

Although we observe no average effect of the MRS explanation, we do note a difference among treated participants, which is explained by a longer time spent on the treatment page. While the treatment is exogenous, we acknowledge that its intensity may not be, which is why Section 3.6.2 further explores potential differences related to individual characteristics. As shown in Figure 3.7a, a large number of participants spent less than 50 seconds on the explanation page. Descriptive statistics reveal that participants spent an average time of 34 seconds reading the MRS explanations. Based on this data, we created a categorical variable to distinguish the intensity of treatment among the treated group. This variable slices the time spent into quartiles and defines the following categories: *No* (*Algorithm* and *Anti-Algorithm* groups), *Low* (≤ 16.25 seconds), *Medium-Low* (> 16.25 , & ≤ 27 seconds), *Medium-High* (> 27 , & ≤ 43.75 seconds), and *High* (> 43.75 seconds). Figure 3.7b illustrates the variation in relative listening time (t_{ik}) across the rounds when grouping by self-exposure to treatment. It can be seen that the *High* treatment group spent almost twice as much time listening to the tracks in the early rounds, compared with the other self-exposure levels, and exceeded the untreated group by 0.1 points. Finally, Figures in Appendix 3.8.3, illustrate the average listening time categorized by track index, revealing distinct patterns across treatment groups. The *High* treatment group demonstrates a consistently higher average listening time compared to other treated groups (i.e., *Low*, *Medium-Low*, *Medium-High*) and to the other two experimental conditions *Anti-Algorithm* and *Algorithm*. However, beyond the 30-track threshold, the average relative listening time levels off and becomes uniformly distributed across all treated groups, except for the *Low* group, as well as across the two experimental conditions. Therefore, we argue that our experimental design allows for a realistic estimation of the treatment effect, as we observe the discovery behaviors over an extended period for which fatigue is easily identifiable and experienced similarly across participants.

In addition to the categorical variable, we also opted for a more direct approach to processing intensity by measuring the time spent on the explanation page. For ease of interpretation, we standardized this time using z-score. As Table 3.7 shows, an increase in self-exposure to treatment leads to an increase in the relative time spent listening (*Model 1*).²¹ The categorical variable further highlights this effect; indeed, the *Low* intensity treatment group shows a lower relative listening time than the untreated groups, 0.027 (SE = 0.010, $p < 0.01$) lower than the control groups. Conversely, participants treated with *High* intensity show a higher relative listening time than the control groups by 0.027 (SE = 0.014, $p < 0.01$, *Model 2*).

To support these results, we use a model that considers possible misunderstandings of the experiment's instructions. Prior to the main part of the experiment, we included questions designed to control the participant's attention to the instructions provided. We excluded from our sample participants who

²¹This model is made within the treated group corresponding to the *Explanations* condition.

Table 3.7: Effect of self-exposure to treatment on Relative Listening

	<i>Dependent variable:</i>		
	(1)	Relative listening (t_{ik}) (2)	(3)
Treatment Time (std.)	0.076** (0.033)		
Anti-Algorithm		Ref.	Ref.
Algorithm		−0.006 (0.009)	−0.004 (0.011)
Algorithm × Low		−0.027*** (0.010)	−0.038*** (0.013)
Algorithm × Medium-Low		−0.014 (0.010)	−0.009 (0.012)
Algorithm × Medium-High		−0.021* (0.012)	−0.017 (0.014)
Algorithm × High		0.027** (0.014)	0.041** (0.018)
Algorithm control	No	Yes	Yes
Participants controls	Yes	Yes	Yes
Round FE	Yes	Yes	Yes
Remove attention check no pass	No	No	Yes
Observations	4,405	11,946	9,122
R ²	0.243	0.243	0.252

Note: Clustered standard error at the participant level in parentheses. Participants' controls include: gender, age, digital PCA, music PCA, discovery mood, and streaming opinion. Track controls include: artist's popularity, track's popularity, track duration, and audio features.

*p<0.1; **p<0.05; ***p<0.01

failed to answer right away to these questions and by such spent the longest time on the attention check page.²² We argue that participants who took longer to answer these questions may not have fully understood the instructions, suggesting they might not have grasped the key principles of the experiment. Results reinforce previous findings, with a noticeable larger effect size for *Low* (−0.038 (SE = 0.013, $p < 0.01$), *Model 3*) and *High* (0.041 (SE = 0.018, $p < 0.01$), *Model 3*) intensity treatment groups. This initial analysis highlights heterogeneity in treatment effects induced by the time of exposure to the MRS explanation.

²²We excluded the participants with an attention check time superior to the 3rd quartile (corresponding to 91.00 seconds). It corresponds to 83 excluded participants (23.08% of the total), and 2,824 observations removed.

3.5.4 Model Predictions

Structural Framework

In our theoretical framework (see Section 3.3), we posit that explanations affect the beliefs about the MRS, shifting these beliefs from θ to θ' . These shifts in beliefs impact the weighting parameters associated with the “consuming” or “exploring” behaviors, denoted respectively by α and β . Specifically, the likelihood of consuming (i.e., consuming = 1) is influenced by the parameters $\alpha(\theta)$ and $\alpha(\theta')$, as described in our theoretical framework (see equations 3.1 and 3.2).

To quantify the effect of the treatment on the propensity to consume, we can express the parameters as follows:

$$\begin{aligned}\alpha(\theta) &= \alpha_0 \\ \alpha(\theta') &= \alpha_0 + \gamma \text{Explanations}\end{aligned}\tag{3.15}$$

where α_0 represents the baseline weighting parameter associated with the “consuming” behavior, and $\gamma \text{Explanations}$ captures the impact of the explanation on the propensity to consume α with updated beliefs θ' .

With this theoretical foundation, we can empirically estimate the propensity to consume by estimating the following Logit model:

$$\mathbb{P}(\text{consume}_{ik} = 1) = \frac{1}{1 + \exp(-(\alpha_0 + \gamma_1 A_k + \gamma_2 (A_k \times E_k) + \gamma_3 X_k + \gamma_4 N_i + \gamma_5 V_{ik}))}\tag{3.16}$$

where α_0 is the baseline propensity to consume, aligning with the theoretical parameter in equation 3.15, A stands for Algorithm, E for the Explanations, X , N and V , are the same vectors of controls outlined in equation 3.14. This structural model connects our theoretical assumptions of belief shifts due to treatment with empirical observations of consuming behavior. The parameter γ_2 in the Logit model directly reflects the influence of treatment on the “consuming” behavior by comparing the *Explanations* group to the *Algorithm* group. When combining γ_1 and γ_2 we can estimate the difference in the propensity to consume between the *Explanations* group and the *Anti-Algorithm* group, similar to the previous models. This approach aligns with our theoretical predictions and provides a framework for empirical analysis, allowing us to quantify the effects of treatment on consuming propensity.

Empirical Results

To test the main hypothesis of our theoretical model on belief updating, and to assess the impact of explanations in more detail (see Section 3.3), we measured participants’ beliefs about the internal workings of the Music Recommendation System (MRS) used in the experiment (either θ or θ'). This measurement

of beliefs was carried out at the end of the experiment, as we chose not to include a prior measurement in order to avoid potential strategic behavior among participants. Moreover, as the experiment was based on a simplified version of a real MRS, measuring initial beliefs could have led to inaccurate assessments of the effects of explanations on belief updating. As a result, we were unable to quantify the magnitude of the belief shock and focused exclusively on an accurate assessment of participants' final beliefs. Participants were asked to indicate the criteria they believed the algorithm used to generate recommendations, which included the following options: recommendations based on stated preferred music genre, stated preferred artists, interactions with the algorithm during the experiment, stated diversity of musical tastes, and listening mood during the experiment.

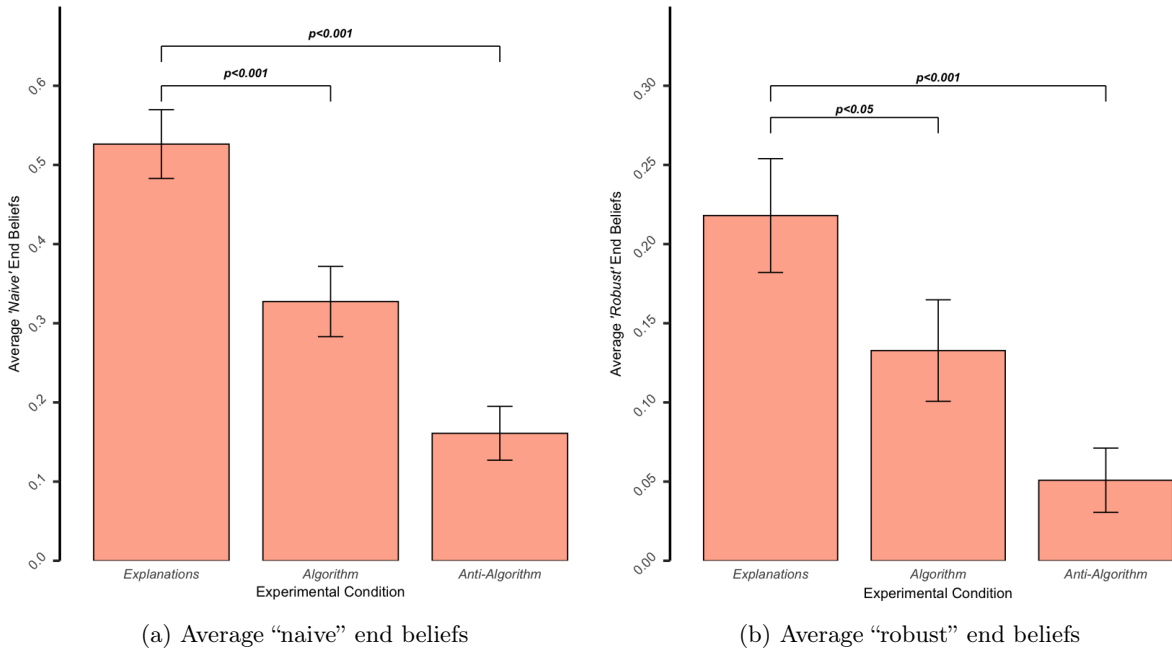


Figure 3.8: Average end beliefs by experimental condition

Figure 3.8a presents the average "naive" end beliefs of participants across different conditions. The "naive" end beliefs measure is assigned a value of one whenever participants include the two key criteria (i.e., the declared preferred genre and diversity of musical tastes) in their end-of-experiment belief answers. The estimates are obtained from OLS regressions ($n = 364$, robust standard errors) of end beliefs on experimental conditions. Compared to the mean end belief score of 0.526 in the *Explanations* condition, our results indicate that participants are significantly more likely to report lower end belief scores in the *Algorithm* condition (0.327; $p < 0.001$) and the *Anti-Algorithm* condition (0.161; $p < 0.001$). In our model, those lower end beliefs are represented by θ , while the higher end beliefs observed in the *Explanations* group are represented by θ' . Figure 3.8b illustrates the average "robust" end beliefs of participants across experimental settings. The "robust" end beliefs measure takes the value of one whenever

a participant only includes the two key criteria in their answers. Estimates are again based on ordinary least squares regressions ($n = 364$, robust standard errors) of end beliefs on experimental conditions. The results presented in Figure 3.8b complement the findings depicted in Figure 3.8b by demonstrating that the average “robust” end beliefs in the *Explanations* setting is consistently significantly higher than in the *Algorithm* and the *Anti-Algorithm* settings.²³ While the exact belief shift before and after the explanation predicted by our theoretical model is not measured, these findings suggest that explanations, on average, contribute to a significant shift in the distribution of individuals’ beliefs regarding the inner workings of MRSs confirming our central theoretical hypothesis that $\theta' > \theta$.

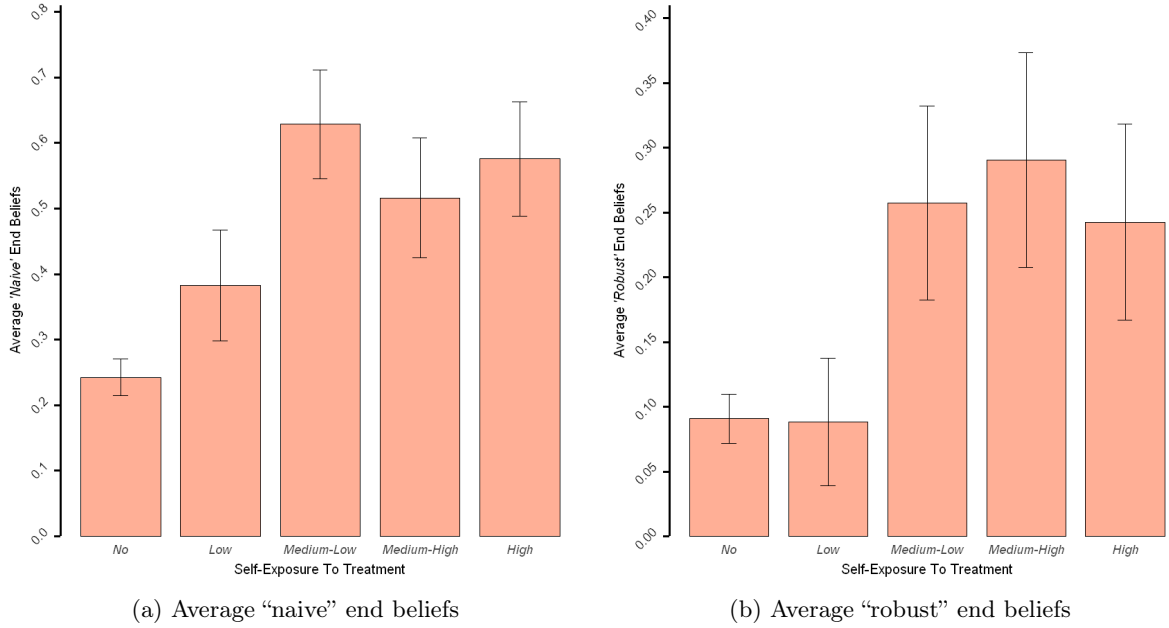


Figure 3.9: Average end beliefs by self-exposure categories

Figures 3.9a and 3.9b display the average “naive” end beliefs and “robust” end beliefs, respectively, across various levels of self-exposure (see Section 3.5.3). Figure 3.9a reveals a positive and significant difference in “naive” end beliefs between each self-exposure category and the two control conditions. However, when examining “robust” end beliefs in Figure 3.9b, there are no longer statistically significant differences between the *No* conditions (*Algorithm* and *Anti-Algorithm*) and the *Low* self-exposure category. Additionally, both figures show that there are no statistically significant differences between the *Medium-Low*, *Medium-High*, and *High* intensity categories, and as such they share comparable belief distributions. These results suggest that self-exposure to treatment, which could be interpreted as a measure of participants’ engagement with the explanations, influences not only discovery behaviors but also

²³Table 3.A.10 shows regression results that control for individual characteristics. While the results align with Figure 3.8, the difference in “robust” end beliefs between the *Explanations* and *Algorithm* conditions becomes less statistically significant (at the 10% level) when these characteristics are considered. This change in significance is mainly due to the variation in participants’ exposure to the treatment, specifically, the time spent on the explanations page, and is further examined in Figure 3.9 and Table 3.A.11.

the distribution of beliefs. This degree of commitment appears to significantly affect how participants form and update their beliefs about the MRS. This finding indicates a complex relationship between beliefs and commitment. Taking into account treatment heterogeneity, we also empirically estimate the propensity to consume by specifying a secondary Logit model:

$$\mathbb{P}(\text{consume}_{ik} = 1) = \frac{1}{1 + \exp(-(\alpha_0 + \gamma_1 A_k + \gamma_2(A_k E_{\text{high},k}) + \dots + \gamma_5(A_k E_{\text{low},k}) + \gamma_6 X_k + \gamma_7 N_i + \gamma_8 V_{ik}))} \quad (3.17)$$

Table 3.8 further highlights the importance of the interaction between beliefs and commitment. First, consistent with the results presented in Section 3.5.2, we find no significant treatment effect at the aggregate level (*Models 1, 3 and 5*) when compared to the control conditions.²⁴ However, the predictions of our theoretical model are mainly validated for participants classified in the *High* self-exposure category. Indeed, in all specifications of our model, belonging to this category significantly increases the probability of adopting consuming behaviors. Specifically, when a track is classified as consumed above the mean, the odds of consuming in the *High* self-exposure group are approximately 65.4% higher (OR = 1.654, *Model 2*) than in the control groups.²⁵ When the consuming behavior is set at the median, the odds are 68.2% higher (OR = 1.766, *Model 4*) in the *High* self-exposure category relative to the control groups. Finally, when the threshold is set at the 30-second level, corresponding to the industry standard to generate a stream,²⁶ The odds are 72.8% higher (OR = 1.728, *Model 6*) in the *High* treatment group than in the control groups. Conversely, although participants in the *Medium-Low* and *Medium-High* self-exposure categories show statistically similar belief distributions with the *High* category ($t(61.093) = -0.426, p = 0.671$; $t(65.889) = -0.138, p = 0.891$, respectively), they exhibit no statistically significant differences in consuming behaviors compared to the control conditions.

3.6 Discussion

3.6.1 Engagement, interactions and music discovery behaviors

Our findings suggest that simply providing explanations of the MRS process can only achieve limited outcomes without a sufficient level of user engagement, reinforcing the distinction, detailed by Anjomshoe et al. (2019), between “Explanation Communication” and “Explanation Reception”. Indeed, the similar distribution of beliefs among participants in the *Medium-Low*, *Medium-High*, and *High* treatment groups (see Section 3.5.4) indicates that enhancing participants’ understanding of the MRS can be effectively

²⁴ $\gamma_2(A_k \times E_k)$ in equation 3.16.

²⁵ $\gamma_1 A_k + \gamma_2(A_k \times E_{\text{high},k})$ in equation 3.17.

²⁶This threshold is, among other criteria, a condition to generate a stream, and therefore perceiving royalties as an artist or a label. See Spotify for artists.

Table 3.8: Effect of treatment and self-exposure on Model Predictions

	<i>Dependent variable:</i>					
	$t_0 = \text{mean}$		Consume $t_0 = \text{median}$		$t_0 = 30\text{s}$	
	(1)	(2)	(3)	(4)	(5)	(6)
Anti-Algorithm	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
Algorithm	-0.111 (0.143)	-0.108 (0.143)	-0.174 (0.147)	-0.170 (0.148)	-0.174 (0.148)	-0.170 (0.148)
Algorithm $\times$ Explanations	-0.068 (0.135)		-0.059 (0.142)		-0.054 (0.139)	
Algorithm $\times$ Low		-0.380* (0.224)		-0.313 (0.214)		-0.380* (0.225)
Algorithm $\times$ Medium-Low		-0.154 (0.200)		-0.149 (0.199)		-0.118 (0.206)
Algorithm $\times$ Medium-High		-0.213 (0.224)		-0.1734 (0.251)		-0.259 (0.242)
Algorithm $\times$ High		0.503*** (0.186)		0.520*** (0.196)		0.547*** (0.185)
Controls + FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	11,946	11,946	11,946	11,946	11,946	11,946
Pseudo R^2	0.197	0.202	0.193	0.197	0.180	0.186

Note: Clustered standard error at the participant level in parentheses. Participants' controls include: gender, age, digital PCA, music PCA, discovery mood, and streaming opinion. Track controls include: artist's popularity, track's popularity, track duration, and audio features.

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

achieved through simplified explanations. However, translating this improved understanding into tangible changes in music discovery behaviors proves more challenging. A key limitation of this study is the lack of clearly identified transmission channels between beliefs and engagement. We hypothesize that a deeper understanding of the MRS processes could shift individuals' discovery moods, thereby leading to greater engagement and fostering more deliberate consuming behaviors. However, this study does not incorporate measures of participants' discovery moods either before or after treatment exposure. Additionally, it is possible that the relative simplicity of our MRS, compared to more sophisticated systems used in the streaming music industry, may have led to reduced interest among participants in the *Explanations* condition. This could explain why, despite similar distributions of beliefs, participants in the *Medium-Low* and *Medium-High* groups did not exhibit comparable music discovery behaviors to those in the *High* group.

Moreover, the external validity of our findings is constrained by the laboratory setting of the experi-

ment. Music discovery behaviors in controlled environments may not fully reflect real-world conditions, where users interact with more complex and dynamic recommendation systems in diverse and less structured contexts. In natural settings, external factors such as mood, social influences, and multitasking behaviors likely play a more significant role in shaping music discovery, making it difficult to generalize the results of this study to broader populations. Therefore, while the lab setting allows for a controlled analysis of the relationship between understanding the MRS and engagement, caution should be exercised in extending these findings to real-world scenarios (further explored in Section 3.6.3).

3.6.2 Treatment effect heterogeneity

Previous results have shown that listening decisions (Section 3.5.3) are influenced by the intensity of treatment. Indeed, we found that the more time participants spent on the MRS explanation page, the more accurate their beliefs about it were, resulting in higher relative listening time and an increased propensity to “consume” (see Table 3.7 and Table 3.8). Conversely, the less treatment participants received, the less accurate their beliefs about the MRS (no difference with untreated groups, see Figure 3.9), and the lower their relative listening time (see Table 3.7). Furthermore, we noted in Section 3.6.1 that the change in beliefs induced by the MRS explanation does not necessarily lead to a change in interactions with the MRS. Indeed, although the *Medium-Low* and *Medium-High* treatment groups share the same beliefs about MRS as the *High* treatment group (see Figure 3.9), these beliefs do not translate into similar behaviors (see Table 3.7 and Table 3.8).

This Section is therefore dedicated to identifying what differentiates all these groups, whether the difference is between the *High* and *Low* treatment groups, or the difference between the *Medium* group (combining the *Medium-Low* and *Medium-High* treatment groups) and the *High* group. Table 3.9 provides descriptive statistics related to these three groups,²⁷ highlighting differences in means for numerical variables, and differences in distribution for categorical variable, between groups. We observe a significant difference between the groups concerning prior IT and digital knowledge (approached by the “Digital PCA” variable, see Section 3.4.3, $p = 0.03$), and concerning the MRS quality stated by the participants after proceeding to the main task ($p = 0.04$). The *Low* treated group displays a higher mean of digital PCA than the other two groups (0.81 vs. -0.45 and -0.31). On the contrary, the *Low* treated group has a better opinion about the MRS quality than *Medium* and *High* groups (3.79 vs. 3.33 and 3.64).

These differences provide insights into the mechanisms underlying engagement and interaction. First, the fact that the *Low* treatment group has higher prior IT knowledge could explain their lower engagement with the MRS explanation. As already mentioned, the relative simplicity of our MRS could lead

²⁷The variables presented in Table 3.9 are the variables that exhibit statistically different results between the groups. Detailed results are available in Table 3.A.12 and Table 3.A.13.

Table 3.9: Difference in mean and distribution of variables between self-exposure groups

		Low		Medium		High		
		Mean		Mean		Mean		p-val
Digital PCA		0.81		−0.45		−0.31		0.03**
Streaming opinion		4.85		5.70		5.45		0.07*
MRS quality		3.79		3.33		3.64		0.04**
		Obs.	%	Obs.	%	Obs.	%	p-val
Gender	Female	25	73.52	35	53.03	26	78.78	0.05*
	Male	9	26.48	30	45.45	6	18.18	
	Other	0	0.00	1	1.52	1	3.04	

Note: Differences in means of numeric variables are estimated by ANOVA, differences between groups of categorical variable are estimated by χ^2 test. *p<0.1; **p<0.05; ***p<0.01

participants with higher expertise to believe that they already understand how the recommendation process works (despite the lack of difference with the untreated groups regarding the end beliefs). Conversely, the *Medium* and *High* treatment groups, with more limited prior knowledge, might need to invest more in explanation (resulting in a difference in behaviors for the *High* treatment group) in order to better understand how the MRS works. In addition, we note a lower appreciation of the MRS for the *Medium* group compared to the *High* group. This difference in appreciation could explain the gaps observed between the beliefs and interactions described in the previous sections. However, it remains difficult to determine whether this gap stems from a difference in prior engagement to the MRS explanation, or simply from less enjoyment experienced while listening to the tracks during the interaction.

Table 3.9 presents additional results, although these are less statistically significant. We observe a generally weaker initial opinion of streaming for the *Low* treatment group, compared to the other two groups ($p = 0.07$), which could explain their lower engagement with MRS explanations. In addition, a gender effect emerges: male participants are proportionally more represented in the *Medium* group than in the other two groups ($p = 0.051$). Taken together, these observations suggest that different types of participants interact differently with MRS explanations and with the MRS itself. These initial results offer clues as to the profiles of people potentially more interested in MRS explanations, and pave the way for future research in this field.

3.6.3 External validity and Welfare perspectives

The population taking part in this experiment was not subject to any particular exclusion criteria (apart from possible hearing problems making it difficult to listen to the tracks). However, it remains different from the population that usually uses music streaming services. Our sample is balanced in terms of age, level of education, and socio-professional status, but shows an over-representation of female

participants (see Tables 3.A.5 and 3.A.6). This composition reflects the general population enrolled at LEEP, where the experiment was conducted. On the other hand, users of online music services are more likely to be young (73% of French people use them, but 95% of those under 25 and only 45% of those over 60), male (75% of French men vs. 71% of French women) and of higher socio-professional status (managers and higher intellectual professions are over-represented by 11 percentage points).²⁸ Although controlled in our econometric analysis, this discrepancy between our sample and actual users could be the subject of specific research by conducting a similar experiment directly on a streaming platform, with a population specifically using this kind of service.

Furthermore, the magnitude of the effect measured also plays a significant role in the context of the music industry. We find that the *Algorithm* group listens to an average of 15% of each recommended song, or an average duration of 29.05 seconds per song (Table 3.5). This is close to the 30-second threshold required to generate a stream on Spotify. In comparison, the average listening time is 26.9 seconds for the *Explanations* group, over 2 seconds less. However, Tables 3.7 and 3.8 clarify this difference: participants in the *High* group increase their relative listening time by 0.027, which corresponds to around 5 seconds more on average than the *Algorithm* group, making sure the average listening duration reaches 30 seconds. They were also 72.8% more likely to reach the 30-second threshold than the control groups. These results show that explanations, when combined with increased engagement, can help generate streams and therefore royalties for artists and labels.

Finally, our experimental design limits musical discovery to a single algorithmically generated playlist: the user's decision is based solely on listening to the tracks in this playlist. In a real-life context, listeners can choose from several playlists, a choice that is partly influenced by the playlists' prominence on the platform (Aguilar & Waldfogel, 2021). It would also be relevant to consider how the platform's market power could be strengthened if it offered explanations for its recommendations. Indeed, the platform is the only one that can offer personalized recommendations on its own service, while third parties can only offer non-personalized editorial playlists, which require little or no explanation on how they recommend tracks.

3.7 Conclusion

This paper measures how MRS explanations affect discovery behaviors on music streaming platforms. The experimental design is based on a 20-minute listening decision task, during which participants interact with personalized recommendations, either aligned or reversed to their preferences. Treatments vary according to the degree of transparency about the recommendation process, in order to assess their

²⁸National Center of Music (CNM). Barometer of music usage in France, October 2023.

influence on music discovery. Participants’ knowledge of songs, likes, dislikes, and skips was tracked. A preliminary questionnaire was used to collect individual preferences and behaviors to calibrate the recommendations and integrate them as control variables in subsequent analyses.

Our theoretical framework models how explanations influence user beliefs and decision-making in the presence of an MRS. Using a Bayesian updating mechanism, we hypothesize that explanations reduce information asymmetry on the recommendation process, leading users to shift from exploring to consuming tracks more carefully. The model predicts that users without explanations will explore more to gather information and reduce the asymmetry, while those provided with explanations will engage more in consuming the recommended tracks. These predictions are empirically tested in the experiment, allowing us to assess the behavioral impact of transparency in MRS recommendations.

Despite the absence of significant mean results for all treated participants, we highlight a differentiation of results according to their commitment to the MRS explanations (measured by time spent reading them). For participants who were highly involved in the explanations (*High* treated group), the combination of explanation and commitment leads to an increase in relative listening time, as well as a reinforcement of “consuming” behavior. This difference in behavior is also reflected in the groups’ beliefs about the MRS: generally speaking, the treated groups (with the exception of the *Low* group) have a better understanding of how music recommendation works than the control groups. However, despite similar measures of beliefs between the *Medium-Low*, *Medium-High* and *High* groups, the first two do not behave in the same way as the last: the *Medium-Low* and *Medium-High* groups show shorter relative listening times and no change in “consuming/exploring” behavior compared with the control groups, while the *High* group sees its relative listening time increase, accompanied by a strengthening of “consuming” behavior. These results reveal a mismatch between behaviors and engagement on the one hand, and beliefs about how recommendations work on the other: understanding how the MRS works doesn’t necessarily lead to changes in behaviors; these changes are mainly observed in individuals most involved in the explanation, aligning with the previous literature on XAI (Anjomshoae et al., 2019).

As self-exposure to treatment is the main vector for differentiating the treatment effect, firstly in terms of belief in the MRS, and secondly in terms of behavioral change. An additional analysis consists of observing any characteristics that differentiate the composition of the intensity groups. The results show that the groups differ according to their prior IT knowledge: the people least involved in the MRS explanations are those who report the highest IT knowledge, while the *High* and *Medium* groups declare themselves to have less knowledge on these topics. We also observe that the overall assessment of the quality of recommendations was slightly higher for the *High* group than for the *medium* groups, which could explain part of the discrepancy between beliefs and behaviors observed in prior results. Finally, less significant results show a possible differentiated effect according to the participants’ gender and their

prior opinion on streaming platforms. These different characteristics can already show that the effects of explanations differ depending on who receives them, and also pave the way for possible future research into which user profiles might benefit most from an MRS explanation.

This study highlights the gap between policy demand (in this case, at the European level) and the actual behavior of streaming service users. As we saw earlier, European regulation has created an obligation for transparency in the functioning of algorithmic content recommendation (via the GDPR, the DSA, or the AI Act). However, this transparency, through the explanations in our experimental design (similar to the explanations displayed on streaming platforms, see Figure 3.A.1), has no average effect on discovery behavior. Such an outcome could be compared with previous findings from the literature on privacy, which shows that, in general, the demand for transparency and the fair use of personal data do not always translate into behavioral change when using the service (Acquisti et al., 2016). Similarly, the literature on law and economics shows that few people take the time to actually read the terms and conditions (Y. Bakos et al., 2014), which is consistent with our findings on exposure to treatment: only the participants who actually read the explanations through self-exposure to treatment were influenced by it in their discovery behaviors.

Generalizing explanations in automated recommendation processes could nevertheless have significant welfare implications by reducing information asymmetry and empowering individuals with a better understanding of AI and algorithmic decisions. The core mechanism we examine, namely the trade-off between “consuming” and “exploring” in the presence of algorithmic recommendations and explanatory information, extends to various domains where users must navigate algorithm-generated choices under conditions of information asymmetry. This transparency could foster greater “AI literacy”, enabling end-users to make more informed choices, trust automated systems, and mitigate potential biases in recommendations. As individuals become more knowledgeable, they are less likely to feel manipulated by opaque algorithms, which enhances consumer welfare by increasing their agency and control over digital interactions. However, increased transparency may also shift the burden of understanding complex systems onto users, potentially widening gaps in technological literacy.

This study also has implications for the various stakeholders in the recorded music industry. First, for artists and record companies, the gain in engagement in music discovery can be beneficial. Indeed, greater commitment (by users for whom explanations have an effect), can create new fan bases. For distributors, the effect could be twofold: the increase in “consuming” behavior could, in the short term, have harmful effects, as the time allocated to listening would be divided between fewer artists from the same distributor’s catalog, which could reduce the royalties received. The long-term effect, however, could still be beneficial, thanks to the increased engagement towards the artist, which could encourage consumers to follow future releases with expectation. Finally, the effect on the platform could be twofold,

depending on its business model: “market-centric” or “user-centric” (Moreau et al., 2022). In the first case, the platform would distribute fewer royalties to rights-holders, because for the same amount of time, users would listen to fewer different artists. In the second case, the platform would distribute more royalties to the rights-holders, because users would listen to fewer artists, but more consistently.

Looking ahead, future experimental research should aim to deepen our understanding of these dynamics by exploring the nuanced relationship between user understanding, commitment, engagement, and discovery behaviors. Additionally, research could focus on employing more sophisticated recommendation systems that better mimic the complexity of real-world algorithms, providing insights into how transparency and explanations interact with more advanced AI systems. Furthermore, understanding the long-term effects of user engagement and its potential to reshape streaming business models could offer valuable perspectives for both industry stakeholders and regulators.

3.8 Appendix

3.8.1 Proofs

Explanation Impact on Beliefs

To fully develop the model of how explanations impact beliefs, we need to incorporate a mechanism that captures the influence of explanations on the agents' belief distributions.

First, we define the Explanation variable as follows, let E represent the explanation provided to the agent. E can be quantified based on its informativeness, clarity, credibility, or a combination of these factors. Assume E is a scalar where higher values indicate more informative or credible explanations.

We suppose that the initial beliefs of agents about the MRS are normally distributed: $\theta \sim \mathcal{N}(\mu, \sigma^2)$. The updated beliefs after receiving explanations are denoted as θ' and we assume that the belief update is driven by the explanation E .

Modeling the Impact of Explanations We assume that the explanations provided to agents impact both the mean (μ) and the variance (σ^2) of their beliefs. The updated beliefs θ' can be modeled using a Bayesian updating framework where explanations act as an exogenous shock.

Explanation Effect on Mean (μ): Let $\Delta\mu(E)$ be the change in the mean of the belief distribution due to the explanation. This can be modeled as:

$$\mu' = \mu + \Delta\mu(E)$$

where $\Delta\mu(E)$ is a function of E .

Explanation Effect on Variance (σ^2): Let $\Delta\sigma^2(E)$ be the change in the variance of the belief distribution due to the explanation. This can be modeled as:

$$\sigma'^2 = \sigma^2 - \Delta\sigma^2(E)$$

where $\Delta\sigma^2(E)$ is a function of E .

The updated beliefs distribution is then given by:

$$\theta' \sim \mathcal{N}(\mu', \sigma'^2)$$

Bayesian Updating Framework In a Bayesian updating framework, the posterior beliefs (θ') are updated based on the prior beliefs (θ) and the new information provided by the explanations (E).

The posterior mean (μ') and variance (σ'^2) are given by:

$$\begin{aligned}\mu' &= \frac{\sigma_y^2 \mu + \sigma^2 \mu_y}{\sigma^2 + \sigma_y^2} \\ \sigma'^2 &= \frac{\sigma^2 \sigma_y^2}{\sigma^2 + \sigma_y^2}\end{aligned}$$

We model the effect of explanations as an exogenous shock to the likelihood's parameters (μ_y and σ_y^2):

- Let $\mu_y = \mu + k_\mu \cdot E$
- Let $\sigma_y^2 = \sigma^2 - k_\sigma \cdot E$

here, k_μ and k_σ are parameters that capture the sensitivity of the explanations on the mean and variance, respectively.

Substituting these into the posterior updates:

1. Posterior Mean:

$$\begin{aligned}\mu' &= \frac{(\sigma^2 - k_\sigma \cdot E)\mu + \sigma^2(\mu + k_\mu \cdot E)}{2\sigma^2 - k_\sigma \cdot E} \\ \iff \mu' &= \mu + \frac{k_\mu \cdot E \cdot \sigma^2}{2\sigma^2 - k_\sigma \cdot E}\end{aligned}$$

2. Posterior Variance:

$$\sigma'^2 = \frac{\sigma^2(\sigma^2 - k_\sigma \cdot E)}{2\sigma^2 - k_\sigma \cdot E}$$

So the updated beliefs θ' follow:

$$\theta' \sim \mathcal{N}\left(\mu + \frac{k_\mu \cdot E \cdot \sigma^2}{2\sigma^2 - k_\sigma \cdot E}, \frac{\sigma^2(\sigma^2 - k_\sigma \cdot E)}{2\sigma^2 - k_\sigma \cdot E}\right)$$

Updated Utility Functions Given the updated beliefs, the utility functions for agents of type 1 and type 2 can be written as:

1. For Agents of Type 1 (No Explanations):

$$U_1(\theta) = \alpha_1(\theta) \cdot x + \beta_1(\theta) \cdot z \text{ subject to } x + z \leq T$$

$$\alpha_1(\theta) = \alpha_0 + k_1 \cdot \theta$$

$$\beta_1(\theta) = \beta_0 - k_2 \cdot \theta$$

2. For Agents of Type 2 (With Explanations):

$$U_2(\theta') = \alpha_2(\theta') \cdot x + \beta_2(\theta') \cdot z \text{ subject to } x + z \leq T$$

$$\alpha_2(\theta') = \alpha_0 + k_1 \cdot \theta'$$

$$\beta_2(\theta') = \beta_0 - k_2 \cdot \theta'$$

with α_0 and β_0 the baseline weighting parameters assigned to the consuming (x) and exploring (z) behaviors in the utility function. These baseline parameters represent the inherent preference or propensity of an agent to engage in these behaviors before accounting for the impact of their beliefs (θ).

Even though the functional forms are similar, the key difference lies in the belief parameters θ and θ' . Type 1 agents do not receive explanations, and their beliefs (θ) are based on initial perceptions and experiences. Type 2 agents receive explanations, leading to updated beliefs (θ'), which are modeled to have different means and variances, capturing the impact of the explanations.

Proof of the Hypotheses

Hypothesis 1. We show that the propensity to explore for type 1 agents, β_1 , is greater than their propensity to consume, α_1 . Given the utility functions for type 1 agents:

$$U_1(\theta) = \alpha_1(\theta) \cdot x + \beta_1(\theta) \cdot z \text{ subject to } x + z \leq T$$

under the constraint $x + z = T$ and our assumptions, the value of z (exploring time) should be greater than the value of x (consuming time).

From the Lagrangian method:

$$\mathcal{L}_1 = \alpha_1(\theta) \cdot x + \beta_1(\theta) \cdot z + \lambda(T - x - z)$$

taking the partial derivatives and setting them to zero:

$$\begin{aligned}\frac{\partial \mathcal{L}_1}{\partial x} &= \alpha_1(\theta) - \lambda = 0 \iff \lambda = \alpha_1(\theta) \\ \frac{\partial \mathcal{L}_1}{\partial z} &= \beta_1(\theta) - \lambda = 0 \iff \lambda = \beta_1(\theta)\end{aligned}$$

since $\lambda = \alpha_1(\theta) = \beta_1(\theta)$ leads to a contradiction because $\alpha_1(\theta) \neq \beta_1(\theta)$, we need to solve for x and z under the constraint $x + z = T$.

Thus:

$$\begin{aligned}\alpha_1(\theta) &= \lambda \\ \beta_1(\theta) &= \lambda\end{aligned}$$

Given that $\alpha_1(\theta) \neq \beta_1(\theta)$, we derive the optimal values using the constraint $x + z = T$:

$$\begin{aligned}x_1^* &= \frac{\beta_1(\theta)}{\alpha_1(\theta) + \beta_1(\theta)} T \\ z_1^* &= \frac{\alpha_1(\theta)}{\alpha_1(\theta) + \beta_1(\theta)} T\end{aligned}$$

however, since $\beta_1(\theta) > \alpha_1(\theta)$, it follows that $z_1^* > x_1^*$. Therefore, type 1 agents should spend more time exploring than consuming.

Given that the expected utility for agents of type 1 is:

$$EU_1 = \sum_{i=1}^n U_1(\theta_i) \cdot \mathbb{P}(\theta_i)$$

plugging the optimal values of x and z into the utility function:

$$U_1(\theta_i) = \alpha_1(\theta_i) \cdot x_1^* + \beta_1(\theta_i) \cdot z_1^*$$

substituting x_1^* and z_1^* :

$$\begin{aligned}U_1(\theta_i) &= \alpha_1(\theta_i) \cdot \frac{\beta_1(\theta_i)}{\alpha_1(\theta_i) + \beta_1(\theta_i)} T + \beta_1(\theta_i) \cdot \frac{\alpha_1(\theta_i)}{\alpha_1(\theta_i) + \beta_1(\theta_i)} T \\ \iff U_1(\theta_i) &= T \left(\frac{\alpha_1(\theta_i)\beta_1(\theta_i) + \beta_1(\theta_i)\alpha_1(\theta_i)}{\alpha_1(\theta_i) + \beta_1(\theta_i)} \right) \\ \iff U_1(\theta_i) &= T \left(\frac{2\alpha_1(\theta_i)\beta_1(\theta_i)}{\alpha_1(\theta_i) + \beta_1(\theta_i)} \right)\end{aligned}$$

Therefore, the expected utility is:

$$EU_1 = T \sum_{i=1}^n \left(\frac{2\alpha_1(\theta_i)\beta_1(\theta_i)}{\alpha_1(\theta_i) + \beta_1(\theta_i)} \right) \cdot \mathbb{P}(\theta_i)$$

Hypothesis 2. We follow the same process as for Hypothesis 1, which leads to the following results:

$$\begin{aligned} x_2^* &= \frac{\beta_2(\theta')}{\alpha_2(\theta') + \beta_2(\theta')} T \\ z_2^* &= \frac{\alpha_2(\theta')}{\alpha_2(\theta') + \beta_2(\theta')} T \end{aligned}$$

since $\alpha_2(\theta') > \beta_2(\theta')$, it follows that $x_2^* > z_2^*$. Therefore, type 2 agents should spend more time consuming than exploring.

The corresponding expected utility is:

$$EU_2 = T \sum_{i=1}^n \left(\frac{2\alpha_2(\theta'_i)\beta_2(\theta'_i)}{\alpha_2(\theta'_i) + \beta_2(\theta'_i)} \right) \cdot \mathbb{P}(\theta'_i)$$

3.8.2 Experimental design appendices

Explanations

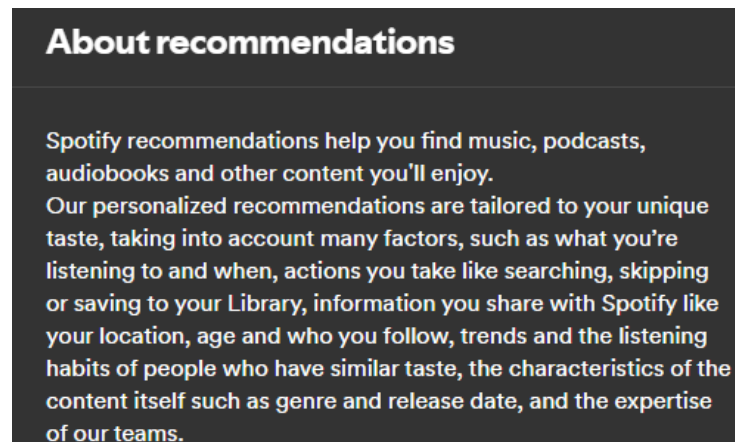


Figure 3.A.1: Spotify's "about recommendations"

Decision Interface

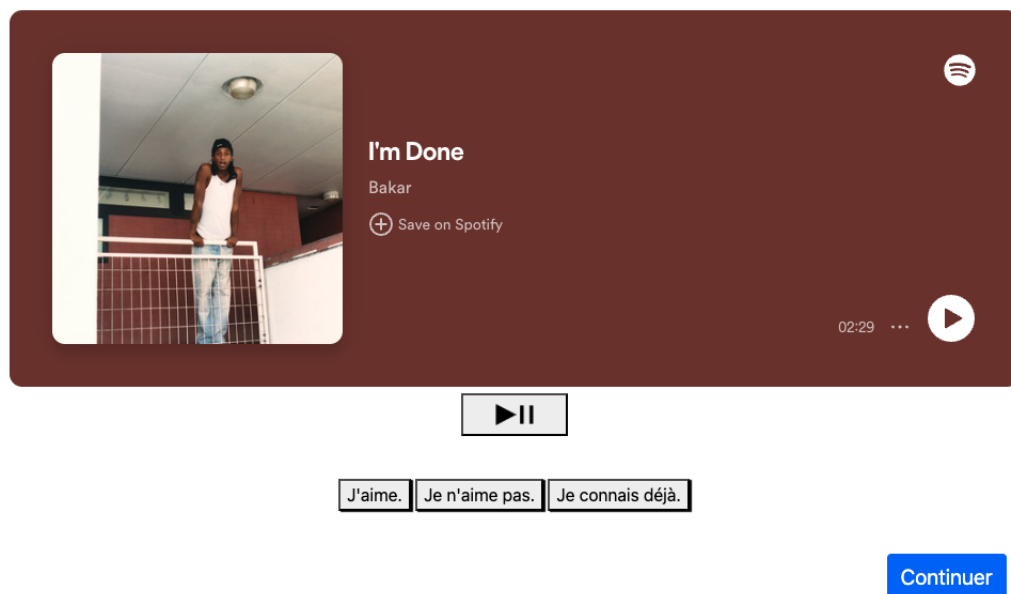


Figure 3.A.2: Decision Interface

Algorithms

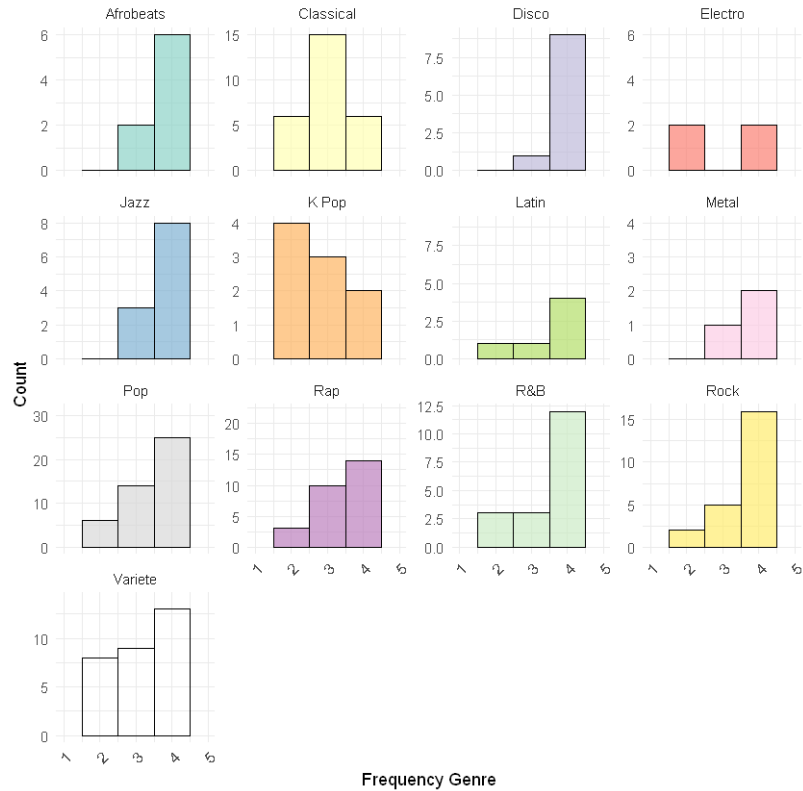


Figure 3.A.3: Distribution of Declared Genre Diversity by Declared Preferred Genre

Table 3.A.1: Results of the Pearson’s Chi-squared Test and Related Statistics

Statistic	Value	<i>p</i> -value
Likelihood Ratio	$\chi^2 = 85.907$, $df = 48$	$p = 6.372e-04$
Pearson	$\chi^2 = 96.648$, $df = 48$	$p = 4.014e-05$
Contingency Coefficient	0.458	
Cramer’s V	0.258	

This indicates a significant association between our two features, with both the Likelihood Ratio and Pearson’s Chi-squared tests showing significant *p*-values, and Cramer’s V suggesting a weak to moderate association.

Algorithm 1 Set Music Genre Recommendations**Input:** Participant's preferred genre (noted p_k) and variety of tastes (noted v_k) $\triangleright v_k \in [1 : 5]$ **Output:** I_{Nk} (set of N recommended tracks for participant k)**Initialize audio features parameters:** Set μ_{jk} and σ_j $\triangleright$ with μ_{jk} the average value of the audio feature j in participant's k preferred genre and σ_j the standard deviation of these average values of audio features j across all genres.**Filter songs by genre:** $PrefGenre_{subset} \leftarrow \{x \in DF \mid x[p_k] = 1\}$ $Pref_{Nk} \leftarrow \{PrefGenre_{subset}[i] \mid i \in S\}$ $\triangleright S$ is a random subset of indices from the set of all possible indices of $Pref_{subset}$ such that $|S| = 50$ $OtherGenres_{subset} \leftarrow \{x \in DF \mid x[p_k] = 0\}$ **Initialize the set of Other Genre Recommendations:** $OtherReco \leftarrow \emptyset$ **if** $v_k = 5$ **then** $Other_{Nk} \leftarrow \{OtherGenres_{subset}[i] \mid i \in S\}$ $\triangleright S$ is a random subset of indices from the set of all possible indices of $OtherGenres_{subset}$ such that $|S| = 50$ **else** $\forall v_k \neq 5$, set $Y = 50$ **Initialize the minimization parameter:** $\lambda \mid \sigma_k$

$$\triangleright \lambda = \begin{cases} 0.1 & \text{if } v_k = 4 \\ 0.05 & \text{if } v_k = 3 \\ 0.025 & \text{if } v_k = 2 \\ 0.01 & \text{if } v_k = 1 \end{cases}$$

while $|OtherReco| < \delta_k \cdot Y$ **do**

$$\triangleright \delta_k = \begin{cases} 24 & \text{if } v_k = 4 \\ 12 & \text{if } v_k = 3 \\ 4 & \text{if } v_k = 2 \\ 1 & \text{if } v_k = 1 \end{cases}$$

for $\forall n \in OtherGenres_{subset}$ **do**Set: $Select \leftarrow \text{True}$ **for** $\forall j$ **do****if** $\mu_{jk} - \lambda \cdot \sigma_j \leq f_{jn} \leq \mu_{jk} + \lambda \cdot \sigma_j$ **then** $\triangleright f_{jn}$ represents an audio feature j of a track n $Select \leftarrow \text{False}$ **break****end if****end for****if** $Select$ is **True** **then** $OtherReco \leftarrow OtherReco \cup \{n\}$ **end if****end for** $\lambda \leftarrow \lambda + \gamma$

$$\triangleright \gamma = \begin{cases} 0.1 & \text{if } v_k = 4 \\ 0.05 & \text{if } v_k = 3 \\ 0.025 & \text{if } v_k = 2 \\ 0.01 & \text{if } v_k = 1 \end{cases}$$

end while**end if** $Other_{Nk} \leftarrow \{OtherReco[i] \mid i \in S\}$ $\triangleright S$ is a random subset of indices from the set of all possible indices of $OtherReco$ such that $|S| = 50$ $I_{Nk} \leftarrow \{Other_{Nk}[i] \mid i \in S\} + \{Pref_{Nk}[i] \mid i \in S'\}$ $\triangleright S$ is a random subset of indices from the set of all possible indices of $Other_{Nk}$ such that $|S| = 25$. The same goes for $Pref_{Nk}$ with $|S'| = 25$. As such, the final length $|I_{Nk}|$ is equal to 50 with additional code ensuring that $\forall i, j \in I_{Nk}, i \neq j \implies I_{Nk}[i] \neq I_{Nk}[j]$

Algorithm 2 Set Reversed Recommendations

Input: Participant's preferred genre (noted p_k) and variety of tastes (noted v_k) $\triangleright v_k \in [1 : 5]$

Output: I_{Nk} (set of N recommended tracks for participant k)

Initialize audio features parameters: Set μ_{jk} and σ_j $\triangleright$ with μ_{jk} the average value of the audio feature j in participant's k preferred genre and σ_j the standard deviation of these average values of audio features j across all genres.

Filter songs by genre:

for $\forall i \neq p_k$ do $\triangleright$ with i as all other genres present in the dataset
 for $\forall j$ do
 Set $\mu_{\text{pref},j}$ and $\mu_{i,j}$ $\triangleright \mu_{\text{pref},j}$ is the average value of audio feature j of tracks in participant's k preferred genre and $\mu_{i,j}$ is the average value of audio feature j for genre i .
 $\Delta_{\text{pref},i,j} = |\mu_{\text{pref},j} - \mu_{i,j}|$
 end for
 $D_{\text{pref},i} = \sum_j \Delta_{\text{pref},i,j}$
 end for

$\bar{p}_k = \underset{i}{\operatorname{argmax}} D_{\text{pref},i}$ $\triangleright$ with $\bar{p}_k$, the furthest away genre from p_k

$\text{RevPrefGenre}_{\text{subset}} \leftarrow \{x \in \text{DF} \mid x[\bar{p}_k] = 1\}$
 $\text{RevPref}_{Nk} \leftarrow \{\text{RevPrefGenre}_{\text{subset}}[i] \mid i \in S\}$ $\triangleright S$ is a random subset of indices from the set of all possible indices of $\text{RevPrefGenre}_{\text{subset}}$ such that $|S| = 50$

$\text{OtherGenres}_{\text{subset}} \leftarrow \{x \in \text{DF} \mid x[p_k] = 0 \text{ and } x[\bar{p}_k] = 0\}$

Initialize Other Genre Recommendations: $\text{OtherReco} \leftarrow \emptyset$

if $v_k = 1$ then
 $\text{Other}_{Nk} \leftarrow \{\text{OtherGenres}_{\text{subset}}[i] \mid i \in S\}$ $\triangleright S$ is a random subset of indices from the set of all possible indices of $\text{OtherGenres}_{\text{subset}}$ such that $|S| = 50$
 else
 $\forall v_k \neq 1$, set $Y = 50$
 Initialize minimized parameter: λ $\triangleright \forall \sigma_k, \lambda = 0.1$
 while $|\text{OtherReco}| < \delta_k \cdot Y$ do
 $\delta_k = \begin{cases} 2.2 & \text{if } v_k = 2 \\ 1.8 & \text{if } v_k = 3 \\ 1.4 & \text{if } v_k = 4 \\ 1 & \text{if } v_k = 5 \end{cases}$
 for $\forall n \in \text{OtherGenres}_{\text{subset}}$ do
 Set: Select $\leftarrow \text{True}$
 for $\forall j$ do
 if $\mu_{jk} - \lambda \cdot \sigma_j \leq f_{jn} \leq \mu_{jk} + \lambda \cdot \sigma_j$ then $\triangleright f_{jn}$ represents an audio feature j of a track n
 Select $\leftarrow \text{False}$
 break $\triangleright$ Exit inner loop if track n doesn't meet criteria
 end if
 end for
 if Select is True then
 $\text{OtherReco} \leftarrow \text{OtherReco} \cup \{n\}$
 end if
 end for
 $\lambda \leftarrow \lambda + \gamma$ $\triangleright \forall \sigma_k, \gamma = 0.1$
 end while
 end if

$\text{Other}_{Nk} \leftarrow \{\text{OtherReco}[i] \mid i \in S\}$ $\triangleright S$ is a random subset of indices from the set of all possible indices of OtherReco such that $|S| = 50$

$I_{Nk} \leftarrow \{\text{Other}_{Nk}[i] \mid i \in S\} + \{\text{RevPref}_{Nk}[i] \mid i \in S'\}$ $\triangleright S$ is a random subset of indices from the set of all possible indices of Other_{Nk} such that $|S| = 25$. The same goes for RevPref_{Nk} with $|S'| = 25$. As such, the final length $|I_{Nk}|$ is equal to 50 with additional code ensuring that $\forall i, j \in I_{Nk}, i \neq j \implies I_{Nk}[i] \neq I_{Nk}[j]$

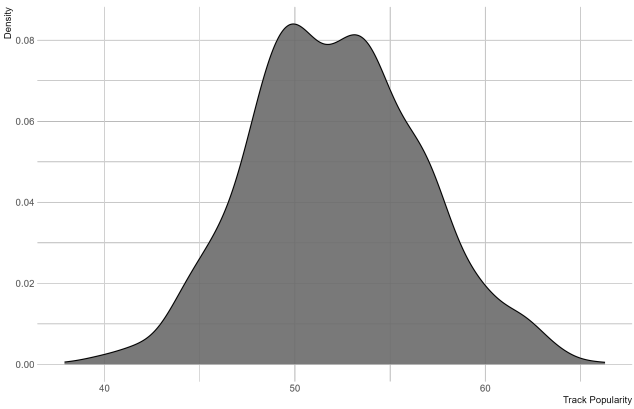


Figure 3.A.4: Density Distribution of the Average Track Popularity per Playlist

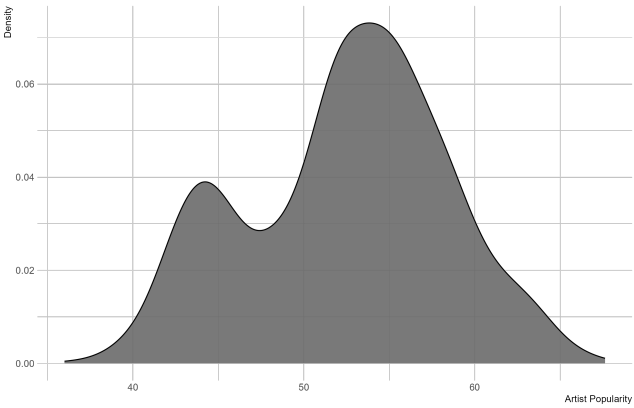


Figure 3.A.5: Density Distribution of the Average Artist Popularity per Playlist

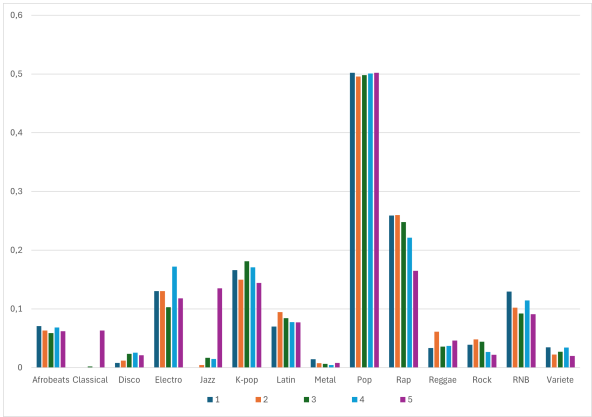


Figure 3.A.6: Genre proportions of generated Pop-oriented playlists

Questionnaire

Table 3.A.2: Preliminary questionnaire

Question	Possible answer(s)
Socio-demographic questions	
What is your age?	<i>Numeric</i>
What is your gender?	Male Female Non-binary Do not prefer to answer
What is your marital status?	Single Married Cohabitation Divorced Widow/Widower Do not prefer to answer
Do you have dependent children?	None 1 child 2 children 3 or more children
What is your socio-professional category?	Farmer Craftsman / Trader / Business manager Executives and Higher intellectual professions Intermediate occupation Employee Worker Student Retired Other
What is your highest level of education obtained?	You have never been to school or you left it before the end of primary school No diploma and schooling interrupted at the end of primary school or before the end of middle school No diploma and schooling until the end of middle school or beyond CEP (primary education certificate) BEPC, Elementary certificate, Middle school certificate, DNB CAP, BEP or equivalent level diploma General or technological baccalaureate, Higher certificate Capacity in law, DAEU, ESEU Professional baccalaureate, Professional technician or teaching certificate, equivalent diploma BTS, DUT, Deug, Deust, Health or social diploma at Bac+2 level, equivalent diploma Bachelor's degree, professional license, Master's degree, equivalent diploma at Bac+3 or Bac+4 level

	Master's degree, DEA, DESS, School diploma at bac+5 level, Health doctorate Research doctorate (excluding health)
Computer and IT knowledge questions	
How often do you connect to the Internet? (At home or outside regardless of the medium)	Every day or almost Several times a week About once a week More rarely Never or almost
How often do you check social media?	Every day or almost Several times a week About once a week More rarely Never or almost
I know how to check that the website on which I'm asked to provide personal data is secure (e.g. https sites, logo or security certificate).	I don't know how to do it I can do it with help I can do it alone I can do this with confidence and, if necessary, guide others
When I'm faced with a technical problem, I'm able to find solutions on the Internet.	I don't know how to do it I can do it with help I can do it alone I can do this with confidence and, if necessary, guide others
I know how to use e-learning tools to improve my digital skills (e.g. video tutorials, online courses).	I don't know how to do it I can do it with help I can do it alone I can do this with confidence and, if necessary, guide others
I know how to protect myself from unwanted and malicious online messages (e.g. spam messages, identity theft e-mails).	I don't know how to do it I can do it with help I can do it alone I can do this with confidence and, if necessary, guide others
When I use a search engine, I can take advantage of its advanced features.	I don't know how to do it I can do it with help I can do it alone I can do this with confidence and, if necessary, guide others
Music knowledge and practices questions	
In general, do you enjoy discovering new songs?	Not at all Very little A little Moderately Medium

Chapter 3 - Algo-Rhythm Unplugged: Effects of explaining algorithmic recommendations on music discovery

	A lot
	Enormously
Have you listened to music in the last 12 months?	Yes, every day or almost Yes, about 3 or 4 days a week Yes, about 1 or 2 days a week Yes, more rarely No, never or almost
Over the past 12 months, what media(s) have you used to listen to music?	CDs or cassettes Vinyls Streaming music on a specialized platform: Spotify, Deezer, Apple Music, etc. Streaming music on another platform (YouTube...) Digital files (MP3 or other) Radio TV
Generally speaking, on what media(s) do you discover new music tracks?	CDs or cassettes Vinyls Streaming music on a specialized platform: Spotify, Deezer, Apple Music, etc. Streaming music on another platform (YouTube...) Digital files (MP3 or other) Radio TV
During your life, what activity(ies) have you practiced?	Playing a musical instrument Singing Dancing Composing on computer None of the activities mentioned
During the last 12 months, have you listened to music in a language other than French?	Yes, often Yes, sometimes No, never or almost
In the last 12 months, have you listened to music on mobile equipment outside your home?	Yes, often Yes, sometimes No, never or almost
Do you ever listen to music for its own sake, without doing anything else?	Yes, every day or almost Yes, quite often Yes, rarely No, never or almost
If you couldn't listen to music anymore, would you miss it?	Yes, a lot Yes, a little No, not so much No, not at all

Do you listen to the radio, whether at home, in the car or elsewhere?	Yes, every day or almost
	Yes, about 3 or 4 days a week
	Yes, about 1 or 2 days a week
	Yes, more rarely
	No, never or almost
Have you attended concerts or cultural events related to music in the last 12 months?	Yes, several times a week
	Yes, about once a week
	Yes, about several times a month
	Yes, Yes, about once every few months
	No, never or almost
If you could no longer attend concerts or cultural events related to music, would you miss it?	Yes, a lot
	Yes, a little
	No, not so much
	No, not at all
Generally speaking, do you enjoy listening to music on [<i>declined according to each support</i>]?	No, not at all
	No
	No, not so much
	Moderately
	Yes, a little
	Yes
	Yes, a lot
Music preferences	
Among the following genres of music, which genre do you prefer (or the one closest to it)?	Afrobeats, Dancehall
	Hip Hop, Rap
	Disco, Funk
	Jazz
	K-Pop, J-Pop
	Latino, Salsa, Reggaeton
	Metal, Hard Rock
	Classical, Opera
	Electronic music, Techno
	Pop, International variety
	RnB, Soul
	Rock
	French variety
Generally speaking, would you say that your tastes in musical genres are:	Exclusive
	Not very varied
	Moderately varied
	Varied
	Very varied
Are there any other musical genres that you particularly like on this list?	Afrobeats, Dancehall

Hip Hop, Rap
 Disco, Funk
 Jazz
 K-Pop, J-Pop
 Latino, Salsa, Reggaeton
 Metal, Hard Rock
 Classical, Opera
 Electronic music, Techno
 Pop, International variety
 RnB, Soul
 Rock
 French variety

Among the following artists, which artist do you prefer (or the one closest to it)?

Rema

 Burna Boy
 Booba
 Eminem
 Bee Gees
 Boney M
 Miles Davis
 Nina Simone
 BTS
 Blackpink
 Metallica
 Iron Maiden
 Mozart
 Beethoven
 David Guetta
 Bob Sinclar
 Taylor Swift
 Dua Lipa
 Usher
 Lauryn Hill
 Nirvana
 Red Hot Chili Peppers
 Céline Dion
 Julien Doré

Are there any other artist(s) that you particularly like on this list?

Rema

 Burna Boy
 Booba
 Eminem
 Bee Gees
 Boney M
 Miles Davis
 Nina Simone
 BTS
 Blackpink
 Metallica
 Iron Maiden

Mozart
 Beethoven
 David Guetta
 Bob Sinclar
 Taylor Swift
 Dua Lipa
 Usher
 Lauryn Hill
 Nirvana
 Red Hot Chili Peppers
 Céline Dion
 Julien Doré

Table 3.A.3: Feedback questionnaire

Feedback questions	
Did you enjoy the songs recommended to you during the experiment?	No, not at all
	No, not so much
	Moderately
	Yes, a little
	Yes, a lot
How well do you understand the music recommendation algorithm?	Not understood at all
	Not understood
	Moderately understood
	Quite well understood
	Very well understood
On what criteria do you think the recommendations were generated?	Your declared favorite musical genre(s)
	Your declared favorite musical artist(s)
	Your interactions with the listening decisions interface
	The diversity of your musical tastes
	Your general mood at the time of listening
How would you characterize the quality of the music recommendation algorithm?	Very bad
	Bad
	Medium
	Good
	Very good
Do you think it would be useful to implement an explanation of how the algorithm works on specialized streaming music platforms using a recommendation algorithm?	No, not useful at all
	No, not very useful
	No, rather not useful
	Moderately useful
	Yes, quite useful

	Yes, useful
	Yes, very useful
MRS beliefs questions	
I think listening to algorithm-generated music recommendations is:	Not useful
	Not very useful
	Rather not useful
	Neutral
	Rather useful
	Useful
	Very useful
How often do you listen to songs recommended by a music recommendation algorithm?	Not at all
	Very little
	A little
	Moderately
	Medium
	A lot
	Enormously
How keen are you on using a music recommendation algorithm to listen to songs right now?	Not at all
	Very little
	A little
	Moderately
	Medium
	A lot
	Enormously
How keen are you on using a music recommendation algorithm to discover songs right now?	Not at all
	Very little
	A little
	Moderately
	Medium
	A lot
	Enormously

Table 3.A.4: Instructions

Thank you for your participation!

You are about to take part in a study of musical discovery. Each participant will receive exactly the same instructions. All decisions and interactions are anonymous.

It is important that you remain silent and do not communicate in any way with other participants outside of your interactions in the experimental interface. If you have any questions, just raise your hand and we'll come and answer them personally.

You will receive a fixed cash payment of 15 euros at the end of the experiment.

Instructions for the main task:

Total listening time You will have to listen to a series of recommended tracks for a total of 20 minutes. At the end of the 20 minutes, you will take the end-of-experience questionnaire, and then receive your remuneration.

Listening to the tracks For each recommended track, you'll have the choice of listening to it for as long as you like. You can scroll through the track as you like using the integrated scroll bar, and pause or play it using the "play/pause" button. You can also "skip" (using the "continue" button) to the next recommendation whenever you like.

Knowledge of the song For each recommended song, you'll be asked to indicate whether or not you know the song. If you already know the song, you'll need to indicate this using the "I already know" button. If you don't click on this button, the song is a discovery for you. *This button is extremely important, as it lets us know whether the recommended track is indeed a musical discovery, so please pay close attention and indicate the tracks you know throughout the main experimental task.*

Like/Dislike For each track, you'll have the opportunity to indicate whether you like it or not. It is not compulsory to indicate your appreciation of the recommendation. If you click on the "I like this song" button, you'll be able to record this song, and a list of all the songs recorded during the experiment will be e-mailed to you (if you wish) at the end of the experiment. *The "I like" and "I don't like" buttons will not influence the recommendation of other tracks you listen to during the experiment.*

SUMMARY OF INSTRUCTIONS

Participants listen to a set of recommended tracks for a total duration of 20 minutes.

You can navigate through the track and move on to the next recommendation whenever you like.

You must indicate whether you know the recommended track by clicking on the "I already know" button; if you don't click, it's a discovery.

You have the option of indicating whether you like the track ("I like this track") or dislike it ("I dislike this track"); this choice will not influence the other tracks recommended. The list of all "liked" songs will be communicated to you at the end of the experiment.

We estimate that the experiment should last around 40 minutes.

Before the main experimental phase, you will be presented with a sample decision interface summarizing all the information.

3.8.3 Additional Results

Additional Figures

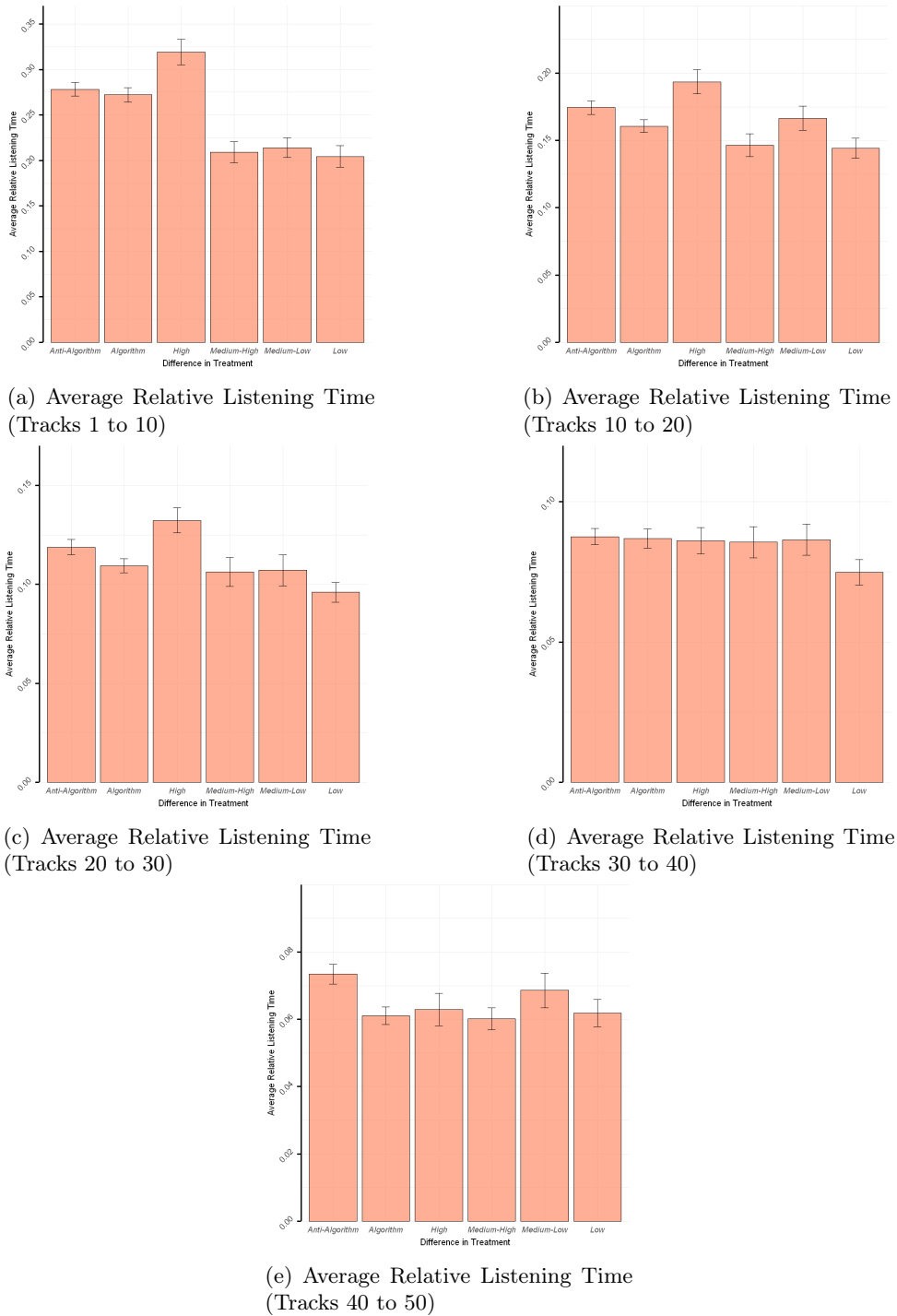
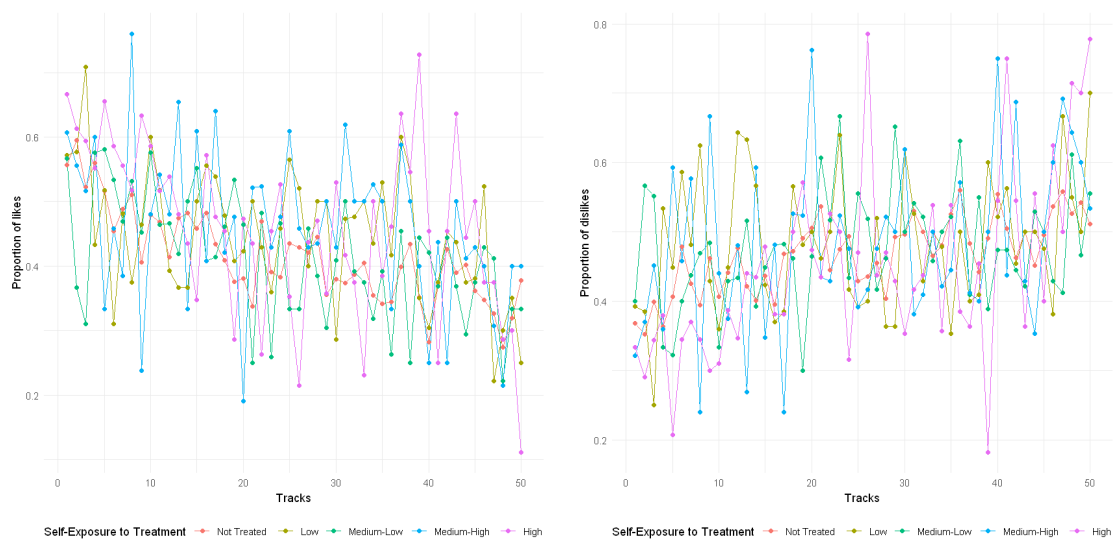


Figure 3.A.7: Average Relative Listening Time by group



(a) Proportion of likes by round and by self-exposure to treatment

(b) Proportion of dislikes by round and by self-exposure to treatment

Figure 3.A.8: Proportion of interactions by round and by self-exposure to treatment

Descriptive statistics of the sample

Table 3.A.5: Summary statistics of numeric variables

	Min	Max	Mean	Median	s.d.	Obs.
<i>Main questionnaire</i>						
Age	18.00	85.00	40.77	36.00	17.05	364
Digital knowledge	13.00	30.00	25.15	25.00	3.66	364
Digital PCA	-3.04	14.42	0.00	-0.10	2.42	364
Music knowledge	11.00	52.00	31.69	32.00	5.64	364
Music PCA	-15.50	16.37	0.00	-0.56	4.58	364
Discovery behavior	1.00	7.00	5.83	6.00	1.19	364
Streaming opinion	1.00	7.00	5.50	6.00	1.74	364
Variety of tastes	1.00	5.00	3.97	4.00	1.02	364
<i>Listening decision</i>						
Number of listened tracks (n)	6.00	50.00	37.59	45.00	14.44	364
Number of known tracks	0.00	45.00	4.77	3.00	6.20	364
Relative listening time (t_{ik})	0.00	1.00	0.15	0.10	0.17	11,946
Number of liked tracks	0.00	50.00	17.31	15.00	10.54	364
Number of disliked tracks	0.00	50.00	15.60	13.00	11.76	364
Number of jumps	0.00	365.00	55.49	40.00	59.86	364
Number of backs	0.00	34.00	1.22	0.00	3.29	364
<i>End questionnaire</i>						
Tracks enjoyment	1.00	5.00	3.64	4.00	0.99	364
MRS understanding	1.00	5.00	3.66	4.00	1.08	364
MRS quality	1.00	5.00	3.47	4.00	0.97	364
Explanation usefulness	1.00	7.00	5.26	5.00	1.37	364
Explanation effect	1.00	5.00	1.94	1.00	1.11	133
Explanation willingness	1.00	5.00	4.05	4.00	0.97	231
Explanation potential effect	1.00	5.00	2.93	3.00	1.25	231
MRS attitude 1	1.00	7.00	4.92	5.00	1.43	364
MRS attitude 2	1.00	7.00	4.91	5.00	1.36	364
MRS behavior	1.00	7.00	4.04	4.00	1.86	364
MRS intention 1	1.00	7.00	4.71	5.00	1.63	364
MRS intention 2	1.00	7.00	4.96	5.00	1.59	364

Table 3.A.6: Summary statistics of categorical variables

		Obs.	%
Treatment allocation	<i>Explanations</i>	133	36.54
	<i>Algorithm</i>	113	31.04
	<i>Anti-Algorithm</i>	118	32.42
Gender	Female	233	64.01
	Male	123	33.79
	Other	8	2.20
Marital status	Single	208	57.14
	Living together	67	18.41
	Married	66	18.13
	Divorced	16	4.39
	Other	7	1.93
Children	None	305	83.80
	1 child	24	6.59
	2 children	17	4.67
	3 children +	18	4.94
Socio-professional status	Craftsman, trader, business manager	15	4.12
	Executive and higher intellectual professions	53	14.56
	Intermediate occupation	31	8.52
	Employee	88	24.17
	Worker	2	< 1.00
	Student	98	26.92
	Retired	50	13.74
	Other	27	7.42
Education	No school or primary education	8	2.20
	Secondary education	93	25.55
	Bachelor's degree or advanced technician's certificate	143	39.28
	Master's degree	114	31.32
	PhD	6	1.65

Table 3.A.7: Balance test between experimental conditions

	p-val	Adjusted p-val
<i>Numeric variables</i>		
Age	0.48	1.00
Digital knowledge	0.95	1.00
Digital PCA	0.83	1.00
Music knowledge	0.86	1.00
Music PCA	0.66	1.00
Discovery behavior	0.32	1.00
Streaming opinion	0.32	1.00
<i>Categorical variables</i>		
Gender	0.84	1.00
Marital status	0.71	1.00
Socio-professional status	0.41	1.00
Education	0.73	1.00

Note: Differences between groups for numeric variables are estimated by ANOVA. Differences between groups for categorical variables are estimated with a χ^2 test. p-val are adjusted with the Bonferroni method. *p<0.1; **p<0.05; ***p<0.01

Treatment effect on engagement metrics

Table 3.A.8: Effect of treatment and self-exposure on Likes and Dislikes

	<i>Dependent variable:</i>			
	Like		Dislike	
	(1)	(2)	(3)	(4)
Anti-Algorithm	Ref.	Ref.	Ref.	Ref.
Algorithm	−0.037 (0.127)	−0.038 (0.127)	0.116 (0.148)	0.115 (0.148)
Algorithm × Explanations	0.077 (0.122)		−0.032 (0.134)	
Algorithm × Low		0.056 (0.173)		0.035 (0.197)
Algorithm × Medium-Low		−0.010 (0.183)		−0.023 (0.191)
Algorithm × Medium-High		0.126 (0.181)		0.021 (0.195)
Algorithm × High		0.167 (0.207)		−0.187 (0.210)
Controls + FE	Yes	Yes	Yes	Yes
Observations	11,946	11,946	11,946	11,946
Pseudo R ²	0.031	0.032	0.017	0.018

Note: Clustered standard error at the participant level in parentheses. Participants' controls include: gender, age, digital PCA, music PCA, discovery mood, and streaming opinion. Track controls include: artist's popularity, track's popularity, track duration, and audio features.

*p<0.1; **p<0.05; ***p<0.01

Table 3.A.9: Effect of treatment and self-exposure on Jumps and Backs

	<i>Dependent variable:</i>			
	Jumps		Backs	
	(1)	(2)	(3)	(4)
Anti-Algorithm	Ref.	Ref.	Ref.	Ref.
Algorithm	−0.156 (0.183)	−0.162 (0.183)	−0.033** (0.014)	−0.033** (0.014)
Algorithm × Explanations	0.053 (0.169)		0.004 (0.008)	
Algorithm × Low		−0.265 (0.221)		0.006 (0.011)
Algorithm × Medium-Low		−0.038 (0.235)		−0.010 (0.007)
Algorithm × Medium-High		0.407 (0.328)		0.005 (0.011)
Algorithm × High		0.202 (0.237)		0.021 (0.017)
Controls + FE	Yes	Yes	Yes	Yes
Observations	11,946	11,946	11,946	11,946
R ²	0.176	0.183	0.014	0.015

Note: Clustered standard error at the participant level in parentheses. Participants' controls include: gender, age, digital PCA, music PCA, discovery mood, and streaming opinion. Track controls include: artist's popularity, track's popularity, track duration, and audio features.

*p<0.1; **p<0.05; ***p<0.01

End Beliefs

Table 3.A.10: Effect of treatment on end beliefs

	<i>Dependent variable:</i>	
	“Naive” End Beliefs	“Robust” End Beliefs
<i>Explanations</i>	Ref.	Ref.
<i>Algorithm</i>	−1.105*** (0.299)	−0.691* (0.370)
<i>Anti-Algorithm</i>	−2.111*** (0.339)	−1.756*** (0.484)
Controls	Yes	Yes
Observations	364	364
Pseudo R ²	0.201	0.113
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01		

Table 3.A.11: Effect of self-exposure to treatment on end beliefs

	<i>Dependent variable:</i>	
	“Naive” End Beliefs	“Robust” End Beliefs
Treated: No	Ref.	Ref.
Treated: Low	1.102** (0.436)	0.180 (0.671)
Treated: Medium-Low	1.989*** (0.440)	1.267*** (0.469)
Treated: Medium-High	1.526*** (0.460)	1.483*** (0.499)
Treated: High	1.574*** (0.436)	1.291*** (0.495)
Controls	Yes	Yes
Observations	364	364
Pseudo R ²	0.187	0.110
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01		

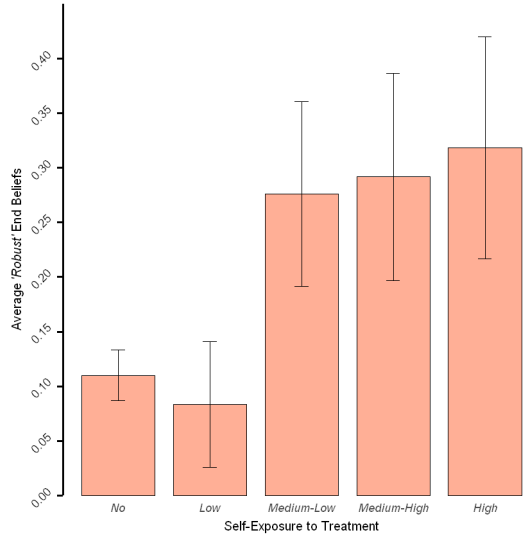


Figure 3.A.9: End Beliefs: Post Attention Check Distribution

Heterogeneity of treatment effect

Table 3.A.12: Difference in mean of numeric variables between self-exposure to treatment groups

	Low	Medium	High	p-val
Age	46.88	40.97	39.79	0.21
Digital knowledge	24.59	25.58	25.00	0.42
Digital PCA	0.81	-0.45	-0.31	0.03**
Music knowledge	31.44	32.53	31.03	0.44
Music PCA	0.76	0.23	-0.29	0.63
Discovery behavior	5.74	5.98	6.09	0.43
Streaming opinion	4.85	5.70	5.45	0.07*
Variety of tastes	3.97	4.18	4.06	0.54
Tracks enjoyment	3.82	3.64	3.85	0.48
MRS understanding	4.00	4.00	3.88	0.79
MRS quality	3.79	3.33	3.64	0.04**

Note:

*p<0.1; **p<0.05; ***p<0.01

Table 3.A.13: Distribution of categorical variables between self-exposure groups

		Low		Medium		High		p-val
		Obs.	%	Obs.	%	Obs.	%	
Gender	Female	25	73.52	35	53.03	26	78.78	0.05*
	Male	9	26.48	30	45.45	6	18.18	
	Other	0	0.00	1	1.52	1	3.04	
Marital status	Single	15	44.12	44	66.66	16	48.48	0.51
	Living together	8	23.53	10	15.15	9	27.27	
	Married	8	23.53	9	13.63	5	15.15	
	Divorced	2	5.88	2	3.03	1	3.03	
	Other	1	2.94	1	1.59	2	6.07	
Socio-professional status	Craftsman, trader, business manager	0	0.00	3	4.54	1	3.04	0.40
	Executive and higher intellectual professions	4	11.76	11	16.66	2	6.06	
	Intermediate occupation	4	11.76	3	4.54	2	6.06	
	Employee	12	35.29	12	18.18	7	21.21	
	Worker	0	0.00	1	1.55	0	0.00	
	Student	6	17.64	19	28.78	12	36.36	
	Retired	7	20.59	14	21.21	5	15.15	
	Other	1	2.96	3	4.54	4	12.12	
	No school or primary education	1	2.95	1	1.53	1	3.04	
Education	Secondary education	14	41.17	16	24.24	8	24.24	0.41
	Bachelor's degree or advanced technician's certificate	12	35.29	26	39.39	17	51.51	
	Master's degree	7	20.59	20	30.30	7	21.21	
	PhD	0	0.00	3	4.54	0	0.00	

Note: *p<0.1; **p<0.05; ***p<0.01

Poisson models

Table 3.A.14: Explanations' effect on Relative Listening (Poisson)

	<i>Dependent variable:</i>		
	(1)	Relative listening (t_{ik}) (2)	(3)
Control groups	Ref.	Ref.	
Explanations	-0.107* (0.058)	-0.091** (0.045)	
Anti-Algorithm			Ref.
Algorithm			-0.032 (0.057)
Algorithm $\times$ Explanations			-0.077 (0.053)
Algorithm control	No	No	Yes
Participants controls	No	Yes	Yes
Round and Track FE	No	Yes	Yes
Observations	11,946	11,946	11,946
Pseudo R^2	0.001	0.056	0.056

Note: Clustered standard error at the participant level in parentheses. Participants' controls include: gender, age, digital PCA, music PCA, discovery mood, and streaming opinion. Track controls include: artist's popularity, track's popularity, track duration, and audio features.

*p<0.1; **p<0.05; ***p<0.01

Table 3.A.15: Effect of self-exposure to treatment on Relative Listening (Poisson)

	<i>Dependent variable:</i>		
	(1)	Relative listening (t_{ik}) (2)	(3)
Treatment Time (std.)	0.484** (0.188)		
Anti-Algorithm		Ref.	Ref.
Algorithm		−0.075 (0.057)	−0.016 (0.067)
Algorithm × Low		−0.211*** (0.080)	−0.301*** (0.100)
Algorithm × Medium-Low		−0.098 (0.073)	−0.063 (0.077)
Algorithm × Medium-High		−0.159* (0.086)	−0.127 (0.010)
Algorithm × High		0.146** (0.070)	0.204** (0.082)
Algorithm control	No	Yes	Yes
Controls + FE	Yes	Yes	Yes
Remove attention check no pass	No	No	Yes
Observations	4,405	11,946	9,122
Pseudo R ²	0.052	0.058	0.061

Note: Clustered standard error at the participant level in parentheses. Participants' controls include: gender, age, digital PCA, music PCA, discovery mood, and streaming opinion. Track controls include: artist's popularity, track's popularity, track duration, and audio features.

*p<0.1; **p<0.05; ***p<0.01

Chapter 4

Chapter 4

Understanding, Trusting, Delegating: How Explanations Shape Public Support for Algorithmic Decision-Making and AI Systems

Co-written with Béatrice Boulu-Reshef

4.1 Introduction

Algorithmic decision-making (ADM) systems are no longer confined to back-office automation. They increasingly act as allocation rules in both markets and public services: they screen job applicants, price insurance, extend (or deny) credit, target advertising, recommend investments, and rank the information economic agents see online. These systems shape efficiency and distribution by influencing who matches with which opportunities, at what price, and under what constraints. Their diffusion therefore raises a basic question for economists and policymakers: when algorithms govern access to high-stakes resources, what makes these systems socially acceptable, and what types of information make people more willing to rely on them?

Europe is a natural setting in which to study this question. The European Union has pursued a distinctive governance strategy that emphasizes fundamental rights, accountability, and transparency in digital markets and public administration. Recent legislation—including the Digital Services Act

(Regulation (EU) 2022/2065) and the EU Artificial Intelligence Act (Regulation (EU) 2024/1689)—places transparency obligations and information provision at the center of the policy response to algorithmic systems. Yet, despite the prominence of “explainability” in European policy debates, credible evidence on whether explanations actually increase public acceptance remains limited and mixed.

From the perspective of information economics, explanations can be understood as a form of disclosure. They may reduce informational asymmetries between system designers and the public, helping individuals form more accurate beliefs about how decisions are made and why outcomes occur (Adadi & Berrada, 2018; Arrieta et al., 2020). At the same time, explanations can reveal complexity, error, or the use of sensitive signals, potentially increasing concerns about manipulation, discrimination, or loss of control. The welfare and governance value of explanation requirements is therefore an empirical question: do explanations build trust and willingness to rely on algorithms, or do they simply make salient new reasons to object?

This paper provides evidence using a large-scale survey experiment in three major EU member states—France, Germany, and Italy. In June 2025, we fielded an online study on broadly representative samples of adults in each country (3,007 respondents in total). Each respondent evaluates six out of twelve ADM and AI-enabled systems spanning five domains that are central to European economic and social life: economic applications (credit scoring, online advertising, algorithmic trading, robo-advisors), health (insurance premium calculation, medical diagnosis, medical assistants/chatbots), hiring (job application screening), policing (predictive policing), and information and culture (recommendation systems, search engines, social media feeds).

The experimental treatment is deliberately designed to mirror the kind of non-technical disclosures that regulators can realistically require. All respondents first receive a neutral description of what a given system is used for. Respondents in the treatment condition additionally receive a short, standardized explanation of how the system works; respondents in the baseline do not. Throughout, we use “explanation” to mean a brief, non-technical description of the system’s functioning (inputs, basic logic, and intended purpose), rather than technical model interpretability or individualized post-hoc rationales. In this sense, the intervention is intentionally light-touch: it adjusts information and framing while remaining easy to ignore and leaving the choice set unchanged, consistent with the behavioral-economics tradition on “nudges” and choice architecture (Thaler & Sunstein, 2008; Beshears & et al., 2020). It also parallels a large applied-micro literature on light-touch informational interventions (e.g., brief letters or reminders), whose effectiveness often depends on whether recipients engage with the information (Hyman, 2020; Cohodes, 2022). Finally, the emphasis on a minimal, modular disclosure is motivated by the scaling challenge highlighted by List’s “voltage effect”: interventions that work in contained settings can lose potency when scaled or transported across contexts (List, 2022, 2023). By fielding the same

explanation module across twelve real-world ADM applications and three EU member states, we provide a first stress test of whether scalable explanations shift beliefs, attitudes, and willingness to delegate.

We then measure outcomes that map naturally to economic questions about adoption. We elicit (i) beliefs that proxy perceived information and experience (knowledge, familiarity, exposure, and self-reported use), (ii) evaluative attitudes along dimensions that feature prominently in both policy debates and the explainable AI literature (transparency, fairness, accountability, privacy, trust, and usefulness) (Castelvecchi, 2016; Cai et al., 2019; Shin, 2021; Weitz et al., 2021), and (iii) willingness to delegate decisions to the system, distinguishing weak delegation (using the system as assistance) from strong delegation (ceding decision rights). By combining between-subject randomization with within-respondent evaluation of multiple systems, we can estimate average effects and map heterogeneity across domains while holding constant individual traits.

Three findings stand out. First, explanations reliably increase perceived understanding. Pooling across systems, treated respondents report higher knowledge, familiarity, exposure, and use by roughly 0.15 to 0.25 points on 7-point scales. Second, explanations generate more modest but systematic improvements in evaluative attitudes, concentrated in perceived fairness, trust, usefulness, and (to a lesser extent) accountability; transparency concerns remain essentially unchanged, and privacy responses are mixed across items. Third, explanations modestly increase willingness to delegate, with effects larger for weak than strong delegation (about 0.09 and 0.06 standard deviations, respectively, in our index-based specifications). These average effects mask substantial domain-level heterogeneity: for some applications, explanations move delegation materially (e.g., weak delegation rises by about 0.27 standard deviations for predictive policing and about 0.20 standard deviations for medical diagnosis), while for others, effects are small and often indistinguishable from zero.

Beyond reduced-form estimates, we use a mechanism-consistent decomposition to clarify how explanations shift delegation. The decomposition is consistent with a sequential structure in which explanations change perceived understanding, which covaries with attitudes, which in turn predict delegation (Baron & Kenny, 1986; Bauer, Preacher, & Gil, 2006; Imai, Keele, & Tingley, 2010). Empirically, the direct effect of explanations on delegation is close to zero; most of the total effect operates through improvements in evaluative attitudes, with beliefs playing a supporting role via a smaller beliefs-to-attitudes pathway.

Our contributions are threefold. First, we provide one of the most comprehensive causal assessments of how brief, policy-relevant explanations affect public acceptance of ADM, spanning a broad set of real-world applications rather than a single domain. Second, by studying three large EU member states under a shared and evolving European regulatory framework, we deliver evidence directly relevant to European debates on transparency mandates and citizen trust in automated decision-making. Third, by jointly measuring beliefs, multidimensional attitudes, and delegation preferences, and by decomposing the

total effect into interpretable pathways, we help discipline discussions of explainability with quantitative magnitudes and mechanism-consistent evidence.

The remainder of the paper is organized as follows. Section 4.2 situates our contribution in the literatures on explainable AI and public acceptance of ADM. Section 4.3 describes the survey experiment and measurement strategy. Section 4.4 reports the main results and the decomposition and presents additional analyses and robustness checks. Section 4.5 concludes with implications for the design and regulation of interpretable algorithmic systems in Europe.

4.2 Related Literature

The increasing complexity and “black-box” nature of many machine learning systems has raised concerns about interpretability for both end-users and developers (Castelvecchi, 2016). These concerns have fueled demands for transparency from stakeholders (Preece et al., 2018), regulators, and citizens more broadly, motivating the emergence of Explainable Artificial Intelligence (XAI) as a field aiming to make AI outputs more understandable to humans (Adadi & Berrada, 2018). Beyond transparency (Cheng et al., 2019), XAI seeks to clarify the rationale behind predictions for users and stakeholders (Cai et al., 2019) and, ultimately, to cultivate trust in AI-driven systems (Shin, 2021; Weitz et al., 2021). A human-centered perspective further emphasizes that explanations should be most informative precisely when outputs appear unintuitive or erroneous (Riedl, 2019).

In parallel, a growing interdisciplinary literature documents that public acceptance of ADM hinges not only on perceived performance, but also on ethical and procedural concerns, notably transparency, fairness, accountability, privacy, and trust. In a large-scale survey of Dutch citizens, Aysolmaz et al. (2023) propose and test a model in which transparency concerns shape perceptions of fairness, accountability, and privacy, which in turn predict trust, perceived usefulness, and ultimately intention to adopt ADM systems; they also show that perceptions vary systematically across application scenarios. Building on this line of work, our study adopts a comparable multi-dimensional measurement framework while moving from correlational acceptance models to causal evidence on the role of explanations.

4.2.1 Explanations, transparency, and trust

A common policy and design response to ADM opacity is to mandate or encourage explanations. However, the theoretical effect of explanations on acceptance is ambiguous: explanations may increase perceived understanding and procedural legitimacy, but can also highlight uncertainty, value-laden choices, or trade-offs that reduce confidence. Following Weitz et al. (2021), it is useful to distinguish (i) an “explainability” strand that develops conceptual and technical frameworks for producing explanations, from

(ii) an “explanations” strand that evaluates the consequences of providing explanations to users across domains.

On the explainability side, contributions propose end-to-end frameworks that separate explanation generation, communication, and reception (Anjomshoae et al., 2019), and distinguish between inherently interpretable models and post-hoc explanation tools for complex models (Arrieta et al., 2020; Beaudouin et al., 2020). A central methodological distinction concerns *local* versus *global* explanations (Speith, 2022). Local methods rationalize a single prediction—for example, LIME builds a simple surrogate model around the neighborhood of an instance to approximate the decision boundary (Ribeiro et al., 2016b). Global methods instead summarize the model’s overall behavior, often via feature-importance measures aggregated across instances (Speith, 2022). Perturbation-based Shapley methods such as SHAP (Lundberg & Lee, 2017; S. B. Cohen et al., 2005) and SAGE (Covert et al., 2021) quantify feature relevance by systematically modifying inputs and tracking prediction changes. In practice, many deployed explanations (including in recommender systems) rely on feature-attribution narratives of the form “the system recommended/assigned Y mainly because of A and B ” (Nunes & Jannach, 2017).

On the explanations side, empirical research in HCI and related fields documents mixed behavioral effects, with outcomes depending on explanation format, task stakes, and user expectations. Early experimental evidence in recommender systems shows that increasing transparency can raise trust and acceptance in some settings (Cramer et al., 2008), and that explanation styles can meaningfully affect how users evaluate recommendations (Herlocker et al., 2000). In social media and advertising contexts, explanations can improve perceived understanding and help users recognize algorithmic bias (Rader et al., 2018), while also inducing “algorithmic disillusionment”—a shift in beliefs about how well an algorithm performs (Eslami et al., 2018). Across studies, explanations often increase objective and self-reported understanding, but do not necessarily translate into higher confidence or trust (Shin, 2021). These findings motivate treating explanations as a potentially modest but policy-relevant intervention whose downstream effects operate through beliefs and attitudes, and whose net impact may vary across domains.

4.2.2 Familiarity, experience, and belief formation

A recurring empirical regularity is that familiarity—knowledge, exposure, and experience—correlates with trust and adoption intentions. Importantly, familiarity may matter both as a cognitive channel (understanding what the system does) and as a preference channel (a taste for the familiar under uncertainty). A complementary lens comes from the information systems and management-science literature on technology adoption. In the Technology Acceptance Model (TAM) and its extensions, adoption is driven by beliefs about a system’s perceived usefulness and perceived ease of use, which shape attitudes and behavioral intentions; in settings involving uncertainty and information asymmetries, trust is often

modeled as a key antecedent of acceptance (Davis, 1989; Venkatesh, Morris, Davis, & Davis, 2003; Gefen, Karahanna, & Straub, 2003). Interpreted through an economic lens, these constructs map naturally to perceived payoffs (usefulness) and perceived frictions or cognitive costs (ease of use), while trust affects the subjective distribution of outcomes and thus the willingness to rely on the system.

In economics, a classic example is “familiarity bias” in portfolio choice: investors disproportionately hold familiar assets (e.g., domestic and local equities) (Huberman, 2001). This broader mechanism provides a useful lens for algorithm adoption: unfamiliar systems may face an additional psychological hurdle beyond objective performance. Moreover, the explanation literature suggests that the efficacy of explanations depends on engagement and users’ propensity to process information (e.g., need for cognition), implying heterogeneity in how explanatory content translates into updated beliefs (Martijn et al., 2022).

Our design explicitly measures belief components such as self-reported knowledge and familiarity, perceived exposure, and self-reported use, and then connects these beliefs to evaluative attitudes and willingness to delegate. This structure allows us to assess whether brief explanations primarily shift perceived understanding (a familiarity-related belief) and whether any resulting changes in delegation plausibly operate through attitudinal channels, while remaining sensitive to heterogeneity by baseline familiarity.

4.2.3 Delegation to algorithms and algorithm aversion in economic settings

A closely related strand studies when individuals *delegate* decisions to algorithms rather than merely expressing favorable attitudes. Work on “algorithm aversion” shows that people may avoid algorithms after observing errors, even when algorithms outperform humans, though aversion can be attenuated when users retain some discretion or control (Dietvorst et al., 2015, 2018). In economics and related fields, delegation and reliance are also studied in organizational and market environments where incentives, monitoring, and trust matter. Cross-country evidence links higher social trust to greater delegation within firms or to advisors (Gur & Bjørnskov, 2017; Gennaioli, Shleifer, & Vishny, 2015), and household-finance evidence highlights how delegation choices depend on perceived competence and context (Holzmeister et al., 2023). Field evidence further suggests that worker compliance with algorithmic recommendations can be shaped by how algorithmic rules are implemented and perceived on the ground (Kawaguchi, 2021). Together, these contributions underscore that acceptance is not only attitudinal: it manifests as a willingness to rely on algorithmic outputs in consequential choices.

4.3 Methodology

4.3.1 Survey Design and Items

We study attitudes toward a broad set of algorithmic decision-making (ADM) systems that individuals may encounter over the course of their lives. The set spans both (i) systems that households use directly in everyday settings (e.g., search engines, recommendation systems, social media feeds) and (ii) systems typically operated by organizations or experts that nonetheless shape household outcomes (e.g., credit scoring, insurance pricing, algorithmic trading, predictive policing). In the latter category, individuals are often affected indirectly through institutional decisions or platform-mediated services rather than through direct use.

Our main research question is whether providing brief, non-technical explanations of how ADM systems work affects respondents’ system-specific beliefs and evaluative attitudes, and whether such changes translate into greater willingness to delegate decisions to algorithms. Because each respondent evaluates multiple systems drawn from different domains, we can examine within-respondent variation across systems while controlling for respondent-level socio-demographic characteristics.

Procedure and measures. Each respondent evaluates multiple algorithmic systems. For each system, respondents in the control condition receive the system name and a short definition describing its purpose. Respondents in the explanation condition receive the same definition plus a short procedural explanation of the system’s inner workings. After this information exposure, we measure four belief-related outcomes: self-reported knowledge (K1), self-reported familiarity (K2), perceived exposure (EXP1), and self-reported use (U1). These items capture respondents’ perceived understanding and engagement with the system immediately following the definition or the explanations (Table 4.A.2).

Respondents’ evaluative attitudes are measured next using a battery adapted from prior work (Aysolmaz et al., 2023). The attitude items cover six dimensions: transparency, fairness, accountability, privacy, trust, and perceived usefulness (Table 4.A.2). We also measure willingness to delegate decisions to algorithms using two items that distinguish between weak delegation (using the system as an assistive tool) and strong delegation (ceding decision rights to the system).

Experimental manipulation. The experiment uses a single between-subject manipulation at the respondent level. Respondents are randomly assigned to either a control condition (definition only) or an explanation condition (definition plus a brief description of how the system operates). Randomization is independent across respondents and balanced by design.

4.3.2 ADM and AI systems

Table 4.1: ADM and AI systems studied in the survey

Domain	System
Economic	Credit scoring Online advertising Algorithmic trading Robo-advisors
Health	Insurance premium calculation Healthcare diagnosis (experts) Medical assistants/chatbots
Hiring	Job application screening
Policing/Justice	Predictive policing
Information & Culture	Recommendation systems Search engines Social media feeds

We study 12 algorithmic decision-making (ADM) and AI-enabled systems spanning five domains: economic applications, health, hiring, policing/justice, and information and culture. This design builds on prior survey-based work that measures perceptions of ADM systems in scenario form (e.g., Aysolmaz et al. (2023)) and extends it by covering a wider set of systems that individuals may encounter directly (as users or decision targets) or indirectly (through institutional or market-level decisions).¹

Table 4.1 summarizes the systems and the short descriptions shown to respondents. The selected systems vary in (i) how directly individuals interact with them (e.g., search engines vs. algorithmic trading) and (ii) how consequential the decisions are perceived to be (e.g., medical diagnosis vs. content recommendation). This breadth helps us examine whether the effect of explanations on beliefs, attitudes, and willingness to delegate is stable across contexts or concentrated in particular domains.

4.3.3 Sample and Data collection

Data were collected online using Dynata’s survey panel.² We recruited respondents in three countries—France, Germany, and Italy—with country-specific quotas to obtain samples that are broadly representative of the adult population in terms of gender, age, and geographic location. The final sample includes 3,007 respondents in total: 1,000 in France, 1,003 in Germany, and 1,004 in Italy. Table 4.A.4 reports descriptive statistics by country.

¹Aysolmaz et al. (2023) study five scenarios: social media recommendation, human resources recruitment, insurance premium identification, medical treatment selection, and tax declaration preparation.

²Dynata applies standard panel quality-assurance procedures designed to mitigate inattentive responding and improve data reliability.

Fieldwork was conducted between June 23, 2025, and June 30, 2025. Respondents provided informed consent before participating. The study protocol was preregistered on AsPredicted (Registration #226,393) and received ethical approval from the PSE Institutional Review Board (IRB #2025-026).

4.4 Results

4.4.1 Descriptive Statistics

This section reports descriptive statistics for the main outcomes. We first characterize respondents’ beliefs about a set of algorithmic systems: self-reported knowledge and familiarity, perceived exposure, and self-reported use. We then summarize attitudes toward these systems across six dimensions: transparency, fairness, accountability, privacy, trust, and perceived usefulness. Finally, we document respondents’ willingness to delegate decisions to algorithms, distinguishing between weak and strong forms of delegation. All outcomes are measured on 7-point Likert scales; we report group means and standard deviations. We compare responses between a baseline condition in which respondents receive only a short definition of the system, and a treatment condition that provides the same definition plus a brief explanation of how the system operates. Because each respondent evaluates multiple systems, the Welch and Wilcoxon tests in this section treat repeated observations as independent and should be read as descriptive. Our main treatment-effect estimates below rely on mixed-effects specifications and a respondent-level (cluster) bootstrap that respects the repeated-measures structure.

We present results in three layers. We begin with descriptive differences between treatment and control for beliefs, attitudes, and delegation, pooled across systems and summarized by system in the Appendix. We then report our main mixed-effects estimates and the mechanism-consistent decomposition that links beliefs, attitudes, and delegation. Finally, we report a set of pre-specified and exploratory heterogeneity analyses that probe when explanations matter more (by baseline system familiarity, political self-placement, and attention to the explanatory content).

Beliefs

Table 4.2: Treated vs Baseline: Welch two-sample t -tests and Wilcoxon rank-sum tests

Variable	Mean (Treated)	Mean (Baseline)	t	df	95% CI (diff)	Sig. (Welch / Wilc.)
Know.	4.260	4.002	9.782	18005	[0.206; 0.310]	*** / ***
Fam.	3.935	3.649	10.771	18015	[0.234; 0.338]	*** / ***
Exp.	3.520	3.093	13.721	17985	[0.366; 0.488]	*** / ***
Use	3.278	2.812	15.101	17953	[0.405; 0.526]	*** / ***
Beliefs (Sum)	14.992	13.556	14.619	18031	[1.244; 1.629]	*** / ***
Beliefs (PCA)	0.103	-0.103	13.931	18040	[0.177; 0.235]	*** / ***

Notes: Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. “Welch” refers to two-sample t -tests with unequal variances; “Wilc.” refers to Wilcoxon rank-sum tests.

Appendix Tables 4.A.5 and 4.A.6 report mean belief scores by system (with standard deviations) for the baseline and treatment conditions, respectively, and Appendix Figures 4.A.1–4.A.4 visualize the corresponding distributions by condition. Focusing on the baseline condition (Table 4.A.5), beliefs vary substantially across application domains. Systems embedded in daily digital life—most notably *Search Engines* and *Social Media*—exhibit the highest baseline levels of familiarity, exposure, and use (e.g., *Search Engines*: Familiarity = 4.46, Exposure = 4.48, Use = 4.31). By contrast, more specialized or institutionally distant systems such as *Predictive Policing* are consistently rated lowest (Knowledge = 3.43, Familiarity = 3.07, Exposure = 2.36, Use = 2.16).

Pooling all system evaluations, the treated group reports systematically higher beliefs than the baseline group (Table 4.2). Relative to the baseline, mean self-reported knowledge increases by 0.26 points (4.260 vs 4.002), familiarity by 0.29 (3.935 vs 3.649), perceived exposure by 0.43 (3.520 vs 3.093), and self-reported use by 0.47 (3.278 vs 2.812). The composite beliefs measures exhibit similarly pronounced shifts (Beliefs Sum: 14.992 vs 13.556; Beliefs PCA: 0.103 vs −0.103). These differences are statistically significant under both Welch two-sample *t*-tests and Wilcoxon rank-sum tests (all $p < 0.01$). Substantively, the estimated effects are precise but moderate in magnitude on a 1–7 scale—typically on the order of 0.25 to 0.50 points for the individual belief components.

Treatment effects also vary across systems. Averaging treatment–control differences across the four belief measures (knowledge, familiarity, exposure, and use), the largest gains are observed for *Online Advertising* and *Recommendation* systems, whereas changes are close to zero for already familiar systems such as *Social Media*. For a subset of systems, estimated effects are negative for specific belief components, consistent with respondents revising downward their perceived exposure or self-reported use once they receive a more detailed explanation of how the system operates.

Attitudes

Attitudinal measures toward algorithmic systems are, on average, mildly favorable in both experimental groups. Aggregating across systems (Appendix Tables 4.A.7–4.A.8), respondents in the baseline condition rate fairness and accountability slightly above the neutral midpoint of the 1–7 scale (PF1 = 4.21, PF2 = 4.18; PA1 = 4.08, PA2 = 3.89). Trust and usefulness are similarly above the midpoint (Tr1 = 4.12, Tr2 = 4.33; PU1 = 4.01, PU2 = 4.05). By contrast, perceived concerns are consistently below the midpoint (TC1 = 3.43, TC2 = 3.29), while privacy concerns are mixed—one item falls below the midpoint, whereas the other lies around it (PC1 = 3.35, PC2 = 4.05). The treated condition exhibits a closely related profile (Appendix Tables 4.A.9–4.A.10), with slightly higher average ratings on fairness, trust, and usefulness (e.g., PF1 = 4.34; Tr2 = 4.49; PU1 = 4.16), while transparency remains the lowest-rated dimension (TC1 = 3.45, TC2 = 3.25).

Table 4.3: Treated vs Baseline: Welch two-sample t -tests and Wilcoxon rank-sum tests

Variable	Mean (Treated)	Mean (Baseline)	t	df	95% CI (diff)	Sig. (Welch / Wilc.)
TC1	3.424	3.426	-0.098	18029	[-0.044; 0.040]	/
TC2	3.237	3.281	-2.063	17985	[-0.086; -0.002]	** / *
PF1	4.325	4.208	5.372	18036	[0.075; 0.160]	*** / ***
PF2	4.308	4.187	5.490	18035	[0.078; 0.164]	*** / ***
PA1	4.168	4.101	2.962	18036	[0.023; 0.111]	*** / ***
PA2	3.949	3.898	2.177	18031	[0.005; 0.099]	** / *
PC1	3.323	3.373	-2.213	18039	[-0.094; -0.006]	** / **
PC2	4.156	4.060	4.155	18040	[0.051; 0.142]	*** / ***
Tr1	4.275	4.130	6.432	18040	[0.101; 0.189]	*** / ***
Tr2	4.461	4.349	5.248	18032	[0.070; 0.155]	*** / ***
PU1	4.120	4.014	4.681	18039	[0.062; 0.151]	*** / ***
PU2	4.149	4.058	3.921	18040	[0.045; 0.136]	*** / ***
Attitudes (Sum)	47.895	47.083	4.841	18027	[0.483; 1.140]	*** / ***
Attitudes (PCA)	0.042	-0.042	5.687	18028	[0.055; 0.114]	*** / ***

Notes: Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. “Welch” refers to two-sample t -tests with unequal variances; “Wilc.” refers to Wilcoxon rank-sum tests.

Providing respondents with an explanation of how the systems operate shifts several attitudinal dimensions in a more favorable direction (Table 4.3). The most consistent gains appear for fairness, trust, and usefulness: both fairness items increase by about 0.12 points (PF1: 4.325 vs 4.208; PF2: 4.308 vs 4.187), and trust and usefulness rise by roughly 0.09–0.15 points (Tr1: 4.275 vs 4.130; Tr2: 4.461 vs 4.349; PU1: 4.120 vs 4.014; PU2: 4.149 vs 4.058). These differences are statistically significant under both the Welch and Wilcoxon tests. Privacy-related responses move in opposite directions across items: PC2 increases (4.156 vs 4.060), while PC1 declines slightly (3.323 vs 3.373), consistent with modest re-weighting rather than uniform shifts in privacy concerns.

By contrast, transparency concerns are largely unaffected. TC1 is essentially unchanged (3.424 vs 3.426), and TC2 decreases only marginally (3.237 vs 3.281), with weaker support under the rank-sum test. At the aggregate level, the overall attitude sum rises from 47.083 in the baseline group to 47.895 in the treated group, and the PCA-based attitude index shifts from -0.042 to 0.042 . Substantively, these effects are precisely estimated but modest relative to the 1–7 scale, typically on the order of one-tenth of a point for individual items.

Willingness to delegate

Table 4.4: Treated vs Baseline: Welch two-sample t -tests and Wilcoxon rank-sum tests

Variable	Mean (Treated)	Mean (Baseline)	t	df	95% CI (diff)	Sig. (Welch / Wilc.)
WTD1 (Weak)	4.21	4.10	4.76	18040	[0.07; 0.16]	*** / ***
WTD2 (Strong)	3.77	3.71	2.06	18034	[0.00; 0.10]	** / *

Notes: Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. “Welch” refers to two-sample t -tests with unequal variances; “Wilc.” refers to Wilcoxon rank-sum tests.

We distinguish between weak delegation—willingness to use the system as an assistive tool (WTD1)—and strong delegation—willingness to cede decision rights and let the system decide (WTD2). Willingness

to delegate is systematically higher under weak than strong delegation in both groups, consistent with greater comfort with advisory or limited delegation than with fully automated decision-making (Figures 4.A.17–4.A.18). Pooling across systems, the treatment increases weak delegation from 4.10 in the baseline group to 4.21 in the treated group (Table 4.4), a difference of 0.11 points (95% CI [0.07; 0.16]) that is strongly significant under both Welch and Wilcoxon tests. For strong delegation, the corresponding gap is smaller—3.71 to 3.77 (CI [0.00; 0.10]), and the evidence is weaker: it is significant under Welch at conventional levels but only marginal under Wilcoxon. Overall, the informational treatment produces precisely estimated but modest changes in delegation preferences, concentrated on the weaker form of delegation.

Heterogeneity across systems mirrors this aggregate pattern. Delegation is highest in domains where algorithmic assistance is perceived as routine or beneficial (e.g., *Search Engines* and *Healthcare Diagnosis*) and lowest for *Online Advertising*, especially under strong delegation (Figures 4.A.17–4.A.18).

4.4.2 Empirical Strategy

Our experiment randomly assigns respondents to either (i) a short definition of each system (control) or (ii) the same definition plus a brief explanation of how the system works (treatment). Each respondent evaluates multiple systems, so the data have a repeated-measures structure: respondent i provides outcomes for system s . We ask two related questions. First, does the explanation affect (i) beliefs, (ii) evaluative attitudes, and (iii) willingness to delegate? Second, to what extent are the observed effects consistent with a simple sequential mechanism?

We posit that brief procedural explanations can influence delegation preferences through the following sequential pathway:

$$\text{Explanation} \rightarrow \text{Beliefs} \rightarrow \text{Attitudes} \rightarrow \text{Delegation}.$$

In our setting, “beliefs” refer to respondents’ system-specific subjective perceptions. In particular, their perceived familiarity, knowledge, use, and exposure to the system were elicited immediately after the vignette (definition with or without explanation) and before the attitude and delegation items. We interpret these measures as respondents’ perceived understanding of the system at the time of elicitation; they may reflect both pre-existing views and any updating induced by the informational content.

A complication is that beliefs and attitudes are elicited after treatment exposure and may therefore lie on the causal path from the explanation to downstream outcomes. Conditioning on beliefs in an attitude regression would amount to controlling for a post-treatment mediator, shifting the estimand away from the total effect of treatment on attitudes toward a direct effect that holds the mediator fixed. For clarity, we therefore proceed in two complementary steps: we first estimate reduced-form (total)

treatment effects on each outcome block, and we then implement a recursive three-equation system as a mechanism-consistent decomposition that quantifies the extent to which treatment effects on delegation are consistent with pathways operating through beliefs and/or attitudes. This approach follows the logic of path analysis (Wright, 1921) and the classic product-of-coefficients decomposition used in mediation research (Baron & Kenny, 1986), adapted to a multilevel setting with repeated measures (Bauer et al., 2006). We treat the decomposition as descriptive of plausible pathways rather than as definitive causal mediation, consistent with modern discussions of identification in mediation analysis (Imai et al., 2010).

Variables and structure We work with standardized (z-scored) versions of three outcome blocks: B_{is}^z (a beliefs index), A_{is}^z (an attitude index), and D_{is}^z (willingness to delegate). The beliefs index aggregates respondents' subjective beliefs about the system (e.g., perceived familiarity, self-assessed knowledge, perceived exposure, and self-reported use). Because each respondent evaluates multiple systems, we include a respondent random intercept ($1 \mid i$) to account for within-respondent dependence. We also include system fixed effects (12 systems) and pre-treatment covariates,

$$X_i = \{\text{Country, Age, Gender, Marital status, Education, Employment status, Income}\}.$$

Step 1: Does treatment change beliefs? We first estimate:

$$B_{is}^z = \alpha_s + a T_{is} + X_i' \pi + u_i + \varepsilon_{is}^B. \quad (4.1)$$

The coefficient a captures the effect of the explanation on beliefs (in standard deviation units).

Step 2: Are attitudes associated with beliefs, and does treatment still matter? Next we estimate:

$$A_{is}^z = \beta_s + c T_{is} + b B_{is}^z + X_i' \lambda + v_i + \varepsilon_{is}^A. \quad (4.2)$$

In this equation, b captures the association between beliefs and attitudes, while c captures any remaining treatment association with attitudes conditional on beliefs. As noted above, we do not interpret c as replacing the reduced-form (total) effect of treatment on attitudes; rather, it is part of the mechanism-consistent decomposition (Baron & Kenny, 1986; Bauer et al., 2006).

Step 3: Does delegation covary with attitudes and beliefs, and is there any remaining direct treatment effect? Finally, we estimate:

$$D_{is}^z = \gamma_s + d T_{is} + f A_{is}^z + e B_{is}^z + X_i' \mu + w_i + \varepsilon_{is}^D. \quad (4.3)$$

Here f captures the association between attitudes and delegation, e captures the association between beliefs and delegation, and d captures any remaining treatment association with delegation conditional on attitudes and beliefs.

Turning the three regressions into a decomposition Let $\hat{a}$ be the estimated coefficient on T from (4.1), $\hat{b}$ and $\hat{c}$ the coefficients on B^z and T from (4.2), and $\hat{d}, \hat{e}, \hat{f}$ the coefficients on T, B^z , and A^z from (4.3). In a linear model, the standard product-of-coefficients logic yields the following components (Baron & Kenny, 1986; Bauer et al., 2006):

$$\text{Direct (not via } A \text{ or } B): \quad \hat{d}, \quad (4.4)$$

$$\text{Indirect via perceived understanding } (T \rightarrow B \rightarrow D): \quad \hat{a}\hat{e}, \quad (4.5)$$

$$\text{Indirect via attitudes } (T \rightarrow A \rightarrow D): \quad \hat{c}\hat{f}, \quad (4.6)$$

$$\text{Indirect via perceived understanding and attitudes } (T \rightarrow B \rightarrow A \rightarrow D): \quad \hat{a}\hat{b}\hat{f}, \quad (4.7)$$

$$\text{Total (sum of components):} \quad \hat{d} + \hat{a}\hat{e} + \hat{c}\hat{f} + \hat{a}\hat{b}\hat{f}. \quad (4.8)$$

Interpretation. In a linear recursive system, the sum of the four components equals the model-implied total effect of T on delegation D ; because T is randomly assigned, this total (reduced-form) effect is causal. By contrast, the allocation of that total across components relies on modeling assumptions and is not interpreted as causal mediation without additional identification conditions.

In linear systems, indirect effects follow the product-of-coefficients rule: a one-unit change in T shifts B by $\hat{a}$, and each unit increase in B shifts D by $\hat{e}$, implying an indirect effect $\hat{a}\hat{e}$; analogous reasoning yields $\hat{c}\hat{f}$ and $\hat{a}\hat{b}\hat{f}$ for longer paths. This decomposition maps directly onto our conceptual sequence: explanations may affect delegation directly, indirectly by increasing perceived understanding, indirectly by shifting attitudes, and indirectly through a sequential pathway from perceived understanding to attitudes to delegation.

Following Imai et al. (2010), this product-of-coefficients decomposition coincides with a causal mediation estimated in linear models under sequential ignorability and an additional no-treatment-mediator-interaction assumption (so that mediation effects do not vary with treatment status). Accordingly, we treat the resulting components as mechanism-consistent accounting terms and emphasize total (reduced-form) treatment effects as our primary causal estimates.

Uncertainty: respondent-level bootstrap Because the indirect effects are products of estimated coefficients, we do not rely on standard closed-form variance formulas. Instead, we obtain uncertainty estimates via a respondent-level bootstrap designed to respect the repeated-measures structure of the data

(Bauer et al., 2006). Specifically, we repeatedly resample respondents with replacement and, for each resampled respondent, retain all associated system-level observations. For each bootstrap replicate, we re-estimate the three mixed-effects models, recompute the direct, indirect, and total components implied by the decomposition, and use the resulting empirical sampling distribution to construct percentile confidence intervals based on the 2.5th and 97.5th percentiles.

Interpretation note The recursive system should be read as an *accounting device* rather than as a stand-alone causal claim about mediation. The reduced-form regressions identify the *total* effect of the explanation on each outcome because treatment is randomly assigned. By contrast, once we include post-treatment variables such as beliefs (B_{is}^z) and attitudes (A_{is}^z) on the right-hand side (e.g., when modeling attitudes as a function of beliefs, or delegation as a function of attitudes and beliefs), the corresponding coefficients do not automatically admit a causal interpretation. The reason is that beliefs and attitudes are not randomized: even though they are influenced by treatment, they may also be jointly shaped by unobserved respondent-specific factors (e.g., general optimism about technology, baseline trust, political orientation, or response styles) that also affect delegation. If such unobserved factors exist, then the associations $B \rightarrow A$ and $A \rightarrow D$ in the system may reflect both causal effects and residual confounding.

Interpreting the product terms (e.g., $\widehat{abf}$) as *causal* indirect effects therefore requires additional assumptions beyond treatment randomization. In the potential-outcomes framework, a common sufficient condition is *sequential ignorability*: conditional on treatment and observed pre-treatment covariates, there are no unobserved confounders of the mediator–outcome relationship (and, in the sequential case, no unobserved confounders of the relationship between earlier and later mediators). These assumptions are strong and cannot be tested directly; modern mediation work emphasizes reporting them transparently and, when possible, assessing sensitivity to violations (Imai et al., 2010).

For this reason, we interpret the decomposition in a deliberately modest way: it quantifies how much of the treatment effect on delegation is consistent with pathways operating through beliefs and attitudes under a linear approximation, and it provides a structured summary that helps connect the empirical results to our theoretical sequence. We continue to treat the reduced-form treatment effects as the primary causal estimates, and we treat the decomposition as complementary evidence about plausible channels.

4.4.3 Treatment Effect

Decomposing the treatment effect on delegation

We examine whether the experimental effect of explanations on willingness to delegate can be summarized by a simple sequential structure, in which explanations shift respondents’ beliefs about a system,

Table 4.5: Decomposition of Treatment Effects on Willingness to Delegate (Sum Indices)

	<i>Dependent variable:</i>	
	Willingness to Delegate	
	WTD1 (Weak) (1)	WTD2 (Strong) (2)
Direct ($\widehat{d}$)	−0.002 (0.010)	−0.015 (0.014)
Indirect via beliefs ($\widehat{ae}$)	0.016*** (0.003)	0.006*** (0.001)
Indirect via attitudes ($\widehat{cf}$)	0.064*** (0.015)	0.057*** (0.013)
Indirect via beliefs → attitudes ($\widehat{abf}$)	0.015*** (0.003)	0.014*** (0.002)
Total	0.094*** (0.020)	0.062*** (0.022)
Observations	18042	18042
Respondents	3007	3007
Bootstrap replications	500	500

Note: Entries report respondent-level cluster bootstrap means with bootstrap standard errors in parentheses (500 replications). Outcomes are standardized (z-scores), so effects are in SD units. The beliefs index (B^z) and the attitude index (A^z) are constructed as *sum scores* of their respective item batteries and then standardized. Stars are based on bootstrap two-sided *p*-values. **p*<0.1; ***p*<0.05; ****p*<0.01

which in turn shape their attitudes and ultimately influence their delegation. Table 4.5 reports the resulting decomposition using sum-based indices for beliefs and attitudes, while Table 4.6 replicates the exercise using PCA-based indices. Throughout, outcomes are standardized, so coefficients are expressed in standard deviation units.

Two patterns stand out. First, the estimated direct component of treatment on delegation is small and statistically indistinguishable from zero for both the weak (WTD1) and strong (WTD2) delegation measures. In contrast, the total treatment effect is positive and precisely estimated for both outcomes: explanations increase WTD1 by 0.094 SD and WTD2 by 0.062 SD when using sum indices (Table 4.5). The same qualitative conclusions are obtained with PCA-based indices (Table 4.6), indicating that the decomposition is not sensitive to the specific index construction.

Second, the increase in delegation is primarily consistent with pathways operating through respondents' attitudes. In the sum-index specification, the indirect component operating through attitudes only ($T \rightarrow A \rightarrow D$) accounts for 0.064 SD in WTD1 and 0.057 SD in WTD2, making it the largest single component. A further, smaller but statistically significant component operates through a sequential path-

Table 4.6: Decomposition of Treatment Effects on Willingness to Delegate (PCA indices)

	<i>Dependent variable:</i>	
	Willingness to Delegate WTD1 (Weak) (1)	WTD2 (Strong) (2)
Direct ($\widehat{d}$)	0.008 (0.010)	-0.015 (0.014)
Indirect via beliefs ($\widehat{ae}$)	0.010*** (0.002)	0.006*** (0.001)
Indirect via attitudes ($\widehat{cf}$)	0.056*** (0.015)	0.057*** (0.013)
Indirect via beliefs $\rightarrow$ attitudes ($\widehat{abf}$)	0.019*** (0.004)	0.014*** (0.002)
Total	0.094*** (0.020)	0.062*** (0.022)
Observations	18042	18042
Respondents	3007	3007
Bootstrap replications	500	500

Note: Entries report respondent-level cluster bootstrap means with bootstrap standard errors in parentheses (500 replications). Outcomes are standardized (z-scores), so effects are in SD units. The beliefs index (B^z) and attitude index (A^z) are constructed using principal component analysis (PCA) and then standardized. Stars are based on bootstrap two-sided p -values. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

way in which treatment shifts beliefs, beliefs shape attitudes, and attitudes in turn predict delegation ($T \rightarrow B \rightarrow A \rightarrow D$): 0.015 SD for WTD1 and 0.014 SD for WTD2. The remaining indirect pathway operating through beliefs without attitudes ($T \rightarrow B \rightarrow D$) is positive but comparatively modest (0.016 SD for WTD1 and 0.006 SD for WTD2). Taken together, these results suggest that explanations increase delegation mainly to the extent that they improve evaluative attitudes toward the systems, with beliefs playing a supporting role by contributing to attitude change.

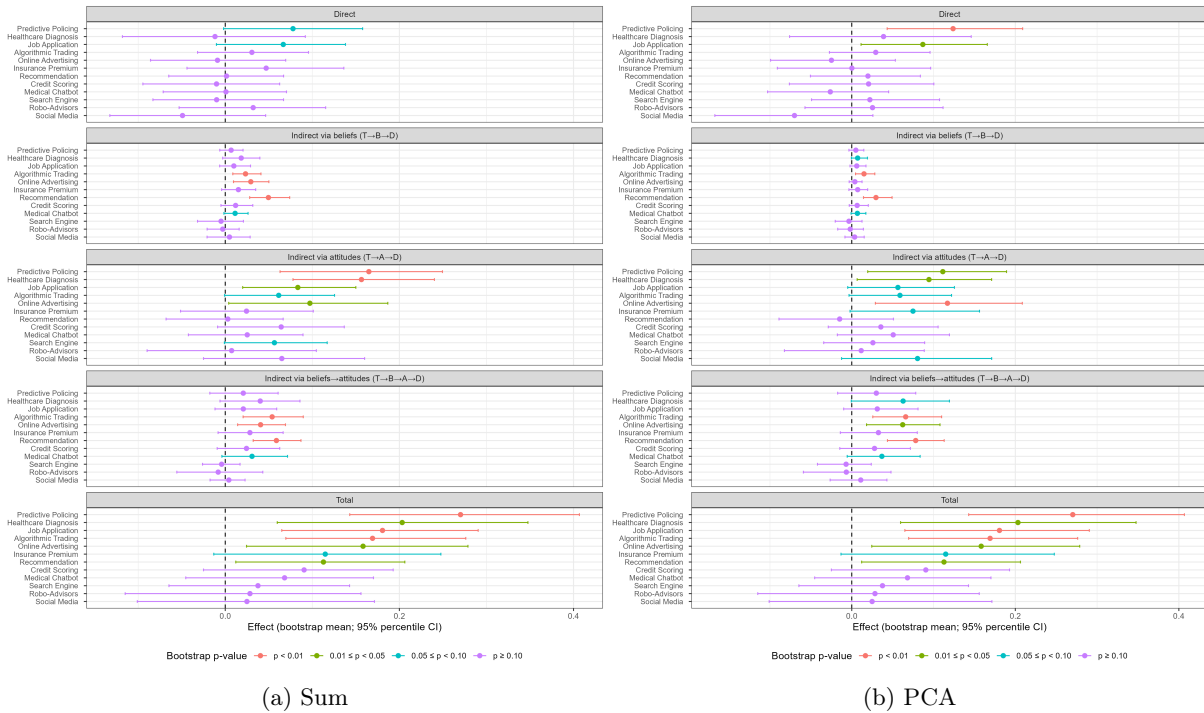
The PCA specification mirrors this pattern closely. The attitudes-only pathway remains the dominant contributor (0.056 SD for WTD1 and 0.057 SD for WTD2), while the sequential beliefs-to-attitudes pathway is of similar magnitude (0.019 SD and 0.014 SD, respectively). The beliefs-only component remains positive but small (0.010 SD for WTD1 and 0.006 SD for WTD2). Across both constructions, therefore, the decomposition implies that explanations primarily matter because they shift evaluative judgments of the systems, rather than because they generate a sizable residual direct propensity to delegate once beliefs and attitudes are taken into account.

Expressed as shares of the total effect (sum-index specification), the attitudes-only pathway $T \rightarrow$

$A \rightarrow D$ accounts for roughly two-thirds of the WTD1 total (0.064 of 0.094) and the vast majority of the WTD2 total (0.057 of 0.062), with the remaining positive contribution split across the beliefs-only and sequential pathways; because the estimated direct component is slightly negative for WTD2, component shares can exceed 100% in absolute value and are therefore interpreted descriptively.

To aid interpretation, we translate standardized effects back to the original 1–7 scale. Using the sample standard deviations ($SD(WTD1) = 1.582$, $SD(WTD2) = 1.676$), the total treatment effect corresponds to an increase of approximately 0.15 points on WTD1 ($0.094 \text{ SD} \times 1.582$) and 0.10 points on WTD2 ($0.062 \text{ SD} \times 1.676$). Relative to the sample means ($\text{mean}(WTD1) = 4.16$, $\text{mean}(WTD2) = 3.74$), these shifts amount to about 3.6% and 2.8%, respectively. The largest component operates through attitudes: the $T \rightarrow A \rightarrow D$ pathway corresponds to roughly 0.10 points on WTD1 and 0.096 points on WTD2 on the original scale. As a diagnostic, we verify that both in the pooled sample and system-by-system, the sum of the four estimated components equals the reduced-form treatment effect of T on delegation (up to numerical tolerance), as implied by the linear recursive system.

Figure 4.1: Decomposition per system, WTD1 (Weak)



Heterogeneity across systems

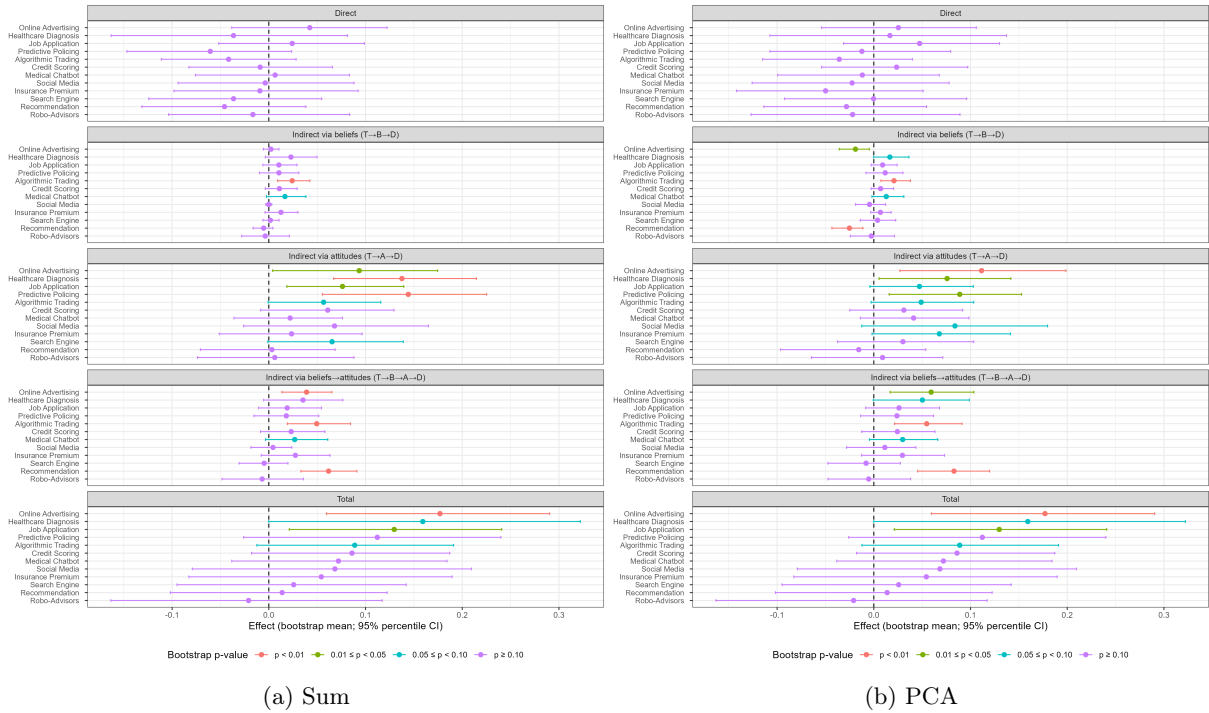
Figures 4.1 and 4.2, together with Appendix Tables 4.A.13 and 4.A.14, show the same decomposition estimated separately for each system; we treat these system-level results as exploratory heterogeneity patterns (without adjusting for multiple comparisons). For WTD1, the explanation increases willingness

to delegate for a subset of systems, with especially large total effects for predictive policing (total ≈ 0.27 SD, about 0.43 points on the 7-point scale) and healthcare diagnosis (total ≈ 0.20 SD, about 0.32 points). Positive and statistically significant totals also appear for algorithmic trading, online advertising, job applications, and recommendation systems (totals roughly between 0.11 and 0.18 SD; about 0.18–0.29 points). For the remaining systems (e.g., robo-advisors, search engines, social media), total effects are smaller and typically not statistically distinguishable from zero.

For WTD2 (strong delegation), the pattern is similar but attenuated: significant total effects are concentrated in fewer systems and are generally smaller in magnitude. The strongest and most robust positive total effects appear for online advertising (total ≈ 0.18 SD, about 0.30 points) and job applications (total ≈ 0.13 SD, about 0.22 points), with healthcare diagnosis also showing a positive total effect (total ≈ 0.16 SD, about 0.27 points). Across many other systems, totals are close to zero and imprecisely estimated.

As a formal test of treatment-effect heterogeneity across the 12 systems, we compared a model with a common treatment effect to a model allowing the treatment effect to vary by system via a full $T \times \text{System}$ interaction. Likelihood-ratio tests show that adding the interaction does not improve fit for either outcome (WTD1: $\chi^2(11)=13.38$, $p=0.27$; WTD2: $\chi^2(11)=10.53$, $p=0.48$). Thus, we find no statistical evidence that the explanation effect on willingness to delegate differs systematically across systems in this specification.

Figure 4.2: Decomposition per system, WTD2 (Strong)



Which components account for system-level effects?

Across systems where the total effect is statistically detectable, the decomposition consistently attributes most of the increase in delegation to attitude-related components. The indirect component operating through attitudes $T \rightarrow A \rightarrow D$ is frequently the largest single term, and it is often complemented by a positive sequential component $T \rightarrow B \rightarrow A \rightarrow D$. By contrast, the direct component is rarely distinguishable from zero at the system level (and when it is, it is typically smaller than the attitude-related terms), consistent with the interpretation that explanations affect delegation primarily by shifting evaluative attitudes rather than by changing delegation “directly.”

Finally, comparing the appendix tables for sum versus PCA constructions shows that the total effects are stable across index choices, while the allocation across belief-based components is somewhat more sensitive to how beliefs are summarized (and can change in magnitude, and occasionally sign, depending on the index construction).

Overall, explanations increase delegation primarily by shifting evaluative attitudes, while changes in perceived understanding matter mainly insofar as they contribute to attitude change. The pooled results establish that explanations increase perceived understanding and improve several evaluative attitudes, with modest downstream effects on delegation. We next unpack (i) which specific items drive these average patterns and (ii) whether effects are systematically stronger in certain domains or for certain respondents.

4.4.4 Heterogeneity and additional analyses

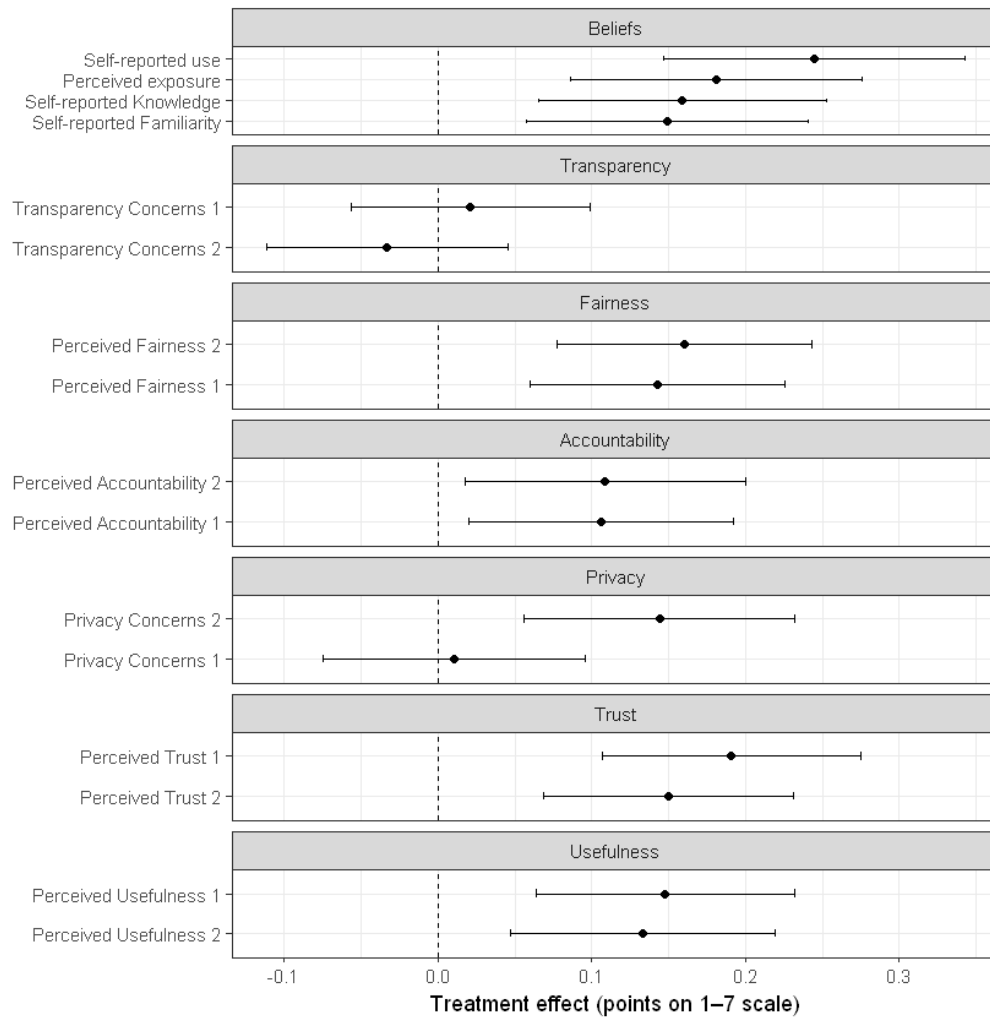
Item-level effects

To complement the index-based results in the main text, Figure 4.3 reports *item-level* treatment marginal gains of providing an explanation. For each belief and attitude item, we estimate a pooled linear mixed model of the form $Y_{is} = \theta T_{is} + \alpha_s + X'_i \pi + u_i + \varepsilon_{is}$, including system fixed effects, the same pre-treatment covariates as in the main specifications, and a respondent random intercept. Coefficients are expressed in points on the original 1–7 response scale. Statistical significance is based on respondent-clustered standard errors.³ Transparency Concerns 1–2 and Privacy Concerns 1 are reverse-coded so that positive coefficients always indicate a shift in a more favorable direction.

Two patterns mirror the main findings. First, explanations generate clear gains in belief-related items. All four belief measures increase significantly, with the largest effect on self-reported use (U1: +0.245 points), followed by perceived exposure (EXP1: +0.181), and the two knowledge and familiarity

³Specifically, inference uses cluster-robust standard errors at the respondent level, so that statistical uncertainty is robust to within-respondent correlation beyond what is captured by the random intercept.

Figure 4.3: Item-level marginal gains of explanations



items (K1: +0.159; K2: +0.149). Thus, explanations substantially increase respondents' perceived understanding and subjective engagement with the systems.

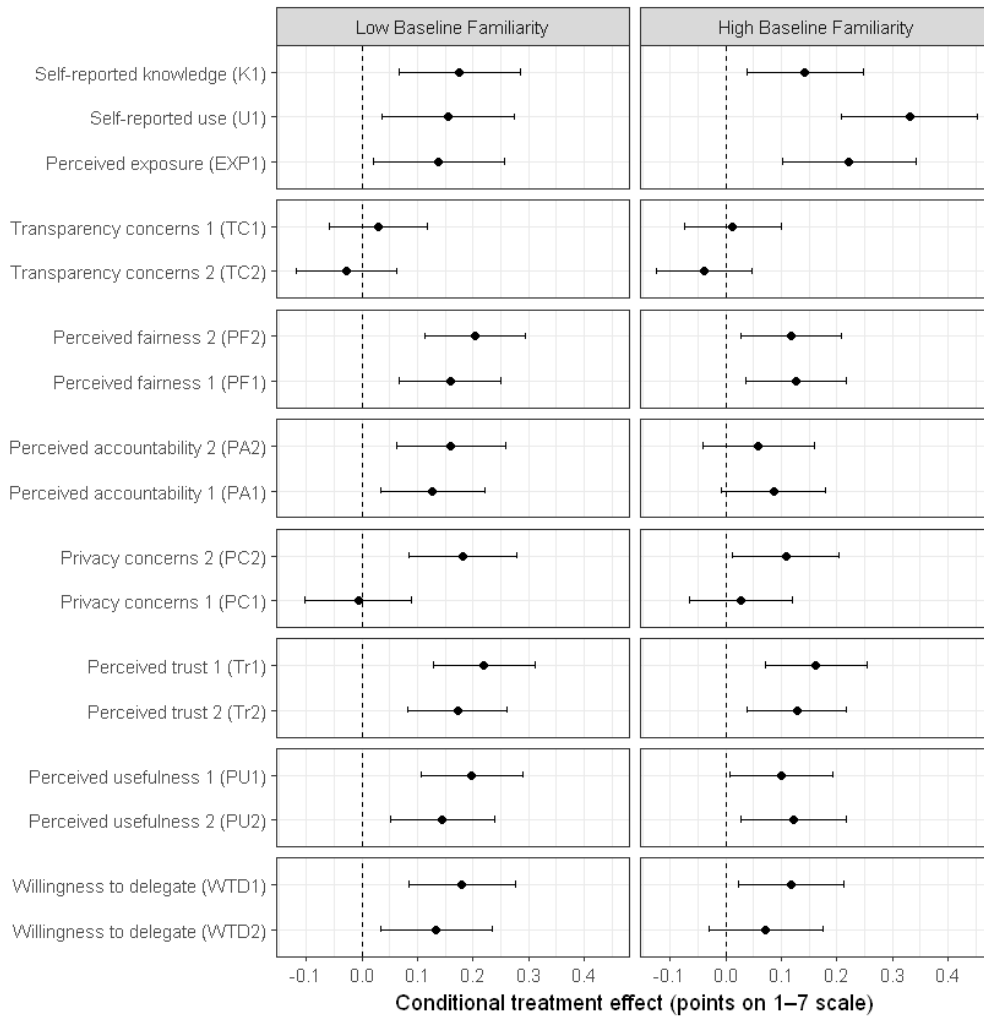
Second, explanations primarily improve evaluative attitudes tied to fairness, trust, accountability, and usefulness, while having a limited impact on transparency and (one dimension of) privacy. Fairness increases by about +0.143 (PF1) to +0.160 (PF2) points; trust increases by +0.150 (Tr2) to +0.191 (Tr1) points; and usefulness increases by +0.133 (PU2) to +0.148 (PU1) points. Accountability also rises modestly (PA1: +0.106; PA2: +0.109). By contrast, the transparency items remain statistically indistinguishable from zero (TC1: +0.021; TC2: -0.033), consistent with the descriptive results showing that explanations do not materially alleviate transparency concerns. For privacy, the “assurance” item (PC2) increases (+0.144), whereas the item capturing concern about personal information use (PC1) shows essentially no change (+0.011).

Overall, the item-level evidence reinforces the interpretation from the decomposition: explanations

reliably increase perceived understanding and improve key evaluative attitudes (especially trust, fairness, and usefulness), while leaving transparency concerns largely unaffected. These patterns help clarify which specific components of the beliefs and attitudes batteries account for the positive, but modest, downstream increase in willingness to delegate reported in the main results.

Heterogeneity by baseline familiarity

Figure 4.4: Conditional marginal gains of explanations by baseline system familiarity.



A natural question is whether explanations generate larger marginal gains for systems that are initially less familiar to respondents (i.e., whether informational returns diminish with baseline familiarity). Because individual self-reported familiarity is measured post-treatment and is itself affected by the explanation, we avoid conditioning on respondents' own familiarity scores. Instead, we use a system-level proxy for baseline familiarity that is computed exclusively in the control group: for each system, we take the mean of the familiarity item (K2) among untreated respondents and standardize this measure across

systems. We then estimate, separately for each survey item, a mixed-effects specification interacting treatment with this baseline familiarity proxy, controlling for system fixed effects, respondent covariates, and a respondent random intercept. Inference uses respondent-clustered standard errors.

In the control group, baseline system familiarity is highest for *Search Engines* ($z = 1.72$), *Social Media* ($z = 1.42$), and *Recommendation* systems ($z = 0.92$), and lowest for *Predictive Policing* ($z = -1.62$) and *Algorithmic Trading* ($z = -1.02$). Systems with above-average baseline familiarity ($z > 0$) include *Search Engines*, *Social Media*, *Recommendation*, *Online Advertising*, and *Insurance Premium*, whereas the remaining systems have below-average baseline familiarity ($z < 0$).

Figure 4.4 reports the implied conditional treatment effects evaluated at low baseline familiarity (-1 SD) and high baseline familiarity ($+1$ SD). Two patterns emerge. First, explanations increase most belief and evaluative attitude items in both contexts. The largest and most consistent gains appear for perceived fairness, accountability, trust, and usefulness: for instance, fairness items increase by about 0.12–0.20 points, trust by about 0.13–0.22 points, and usefulness by about 0.10–0.20 points, depending on the baseline familiarity level. By contrast, transparency items remain close to zero (TC1 and TC2 both have confidence intervals spanning zero at both low and high familiarity), and privacy concerns show mixed results: PC2 increases modestly and positively, while PC1 remains near zero.

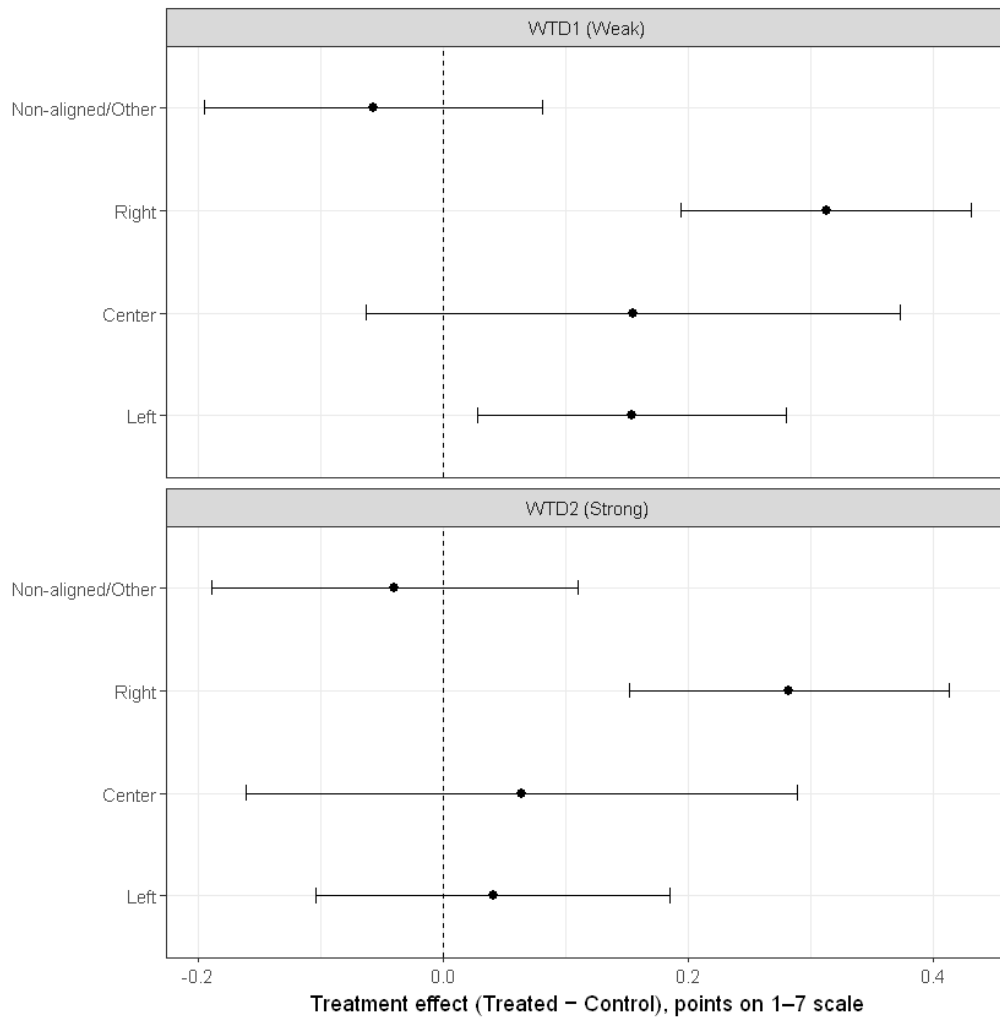
Second, while the point estimates sometimes differ between low and high baseline familiarity, the evidence for systematic moderation is limited. For most items, the conditional effects at -1 SD and $+1$ SD are of similar magnitude and overlap substantially. This is also true for delegation outcomes: both WTD1 and WTD2 are positive at low and high familiarity, but the differences across familiarity levels are not large relative to uncertainty (WTD1: 0.18 vs. 0.12; WTD2: 0.13 vs. 0.07, with the high-familiarity WTD2 interval including zero).

Overall, these exploratory results suggest broadly positive marginal gains of explanations across beliefs and several attitude dimensions, with little clear evidence that returns systematically diminish (or increase) with baseline system familiarity. Where differences appear, they tend to be modest and item-specific rather than a general pattern across constructs.

Heterogeneity by political preferences

Explanations may be processed through broader worldviews about technology, institutions, and regulation. We therefore explore whether the treatment effect differs by respondents' political self-placement, as a parsimonious proxy for such latent predispositions. The political self-placement item (S10) is collapsed into four groups: *Left* (Left-wing/Progressive, Center-left, Green/Environmentalist), *Center* (Centrist/Moderate), *Right* (Center-right, Right-wing/Conservative, Nationalist/Populist), and *Non-aligned/Other* (Libertarian, Other, None/Not affiliated, Prefer not to say). The resulting respondent

Figure 4.5: Treatment effects by political group



counts are: Left ($n = 863$), Center ($n = 280$), Right ($n = 1024$), and Non-aligned/Other ($n = 840$). Treatment assignment is balanced within groups.

We estimate mixed-effects models of willingness to delegate (WTD1 and WTD2) with respondent random intercepts, system fixed effects, and the standard demographic controls, allowing the treatment effect to vary by political group via a treatment-by-group interaction. Likelihood-ratio tests indicate no detectable heterogeneity for weak delegation (WTD1: $\chi^2(3) = 4.06$, $p = 0.255$), but evidence of heterogeneity for strong delegation (WTD2: $\chi^2(3) = 10.85$, $p = 0.013$).

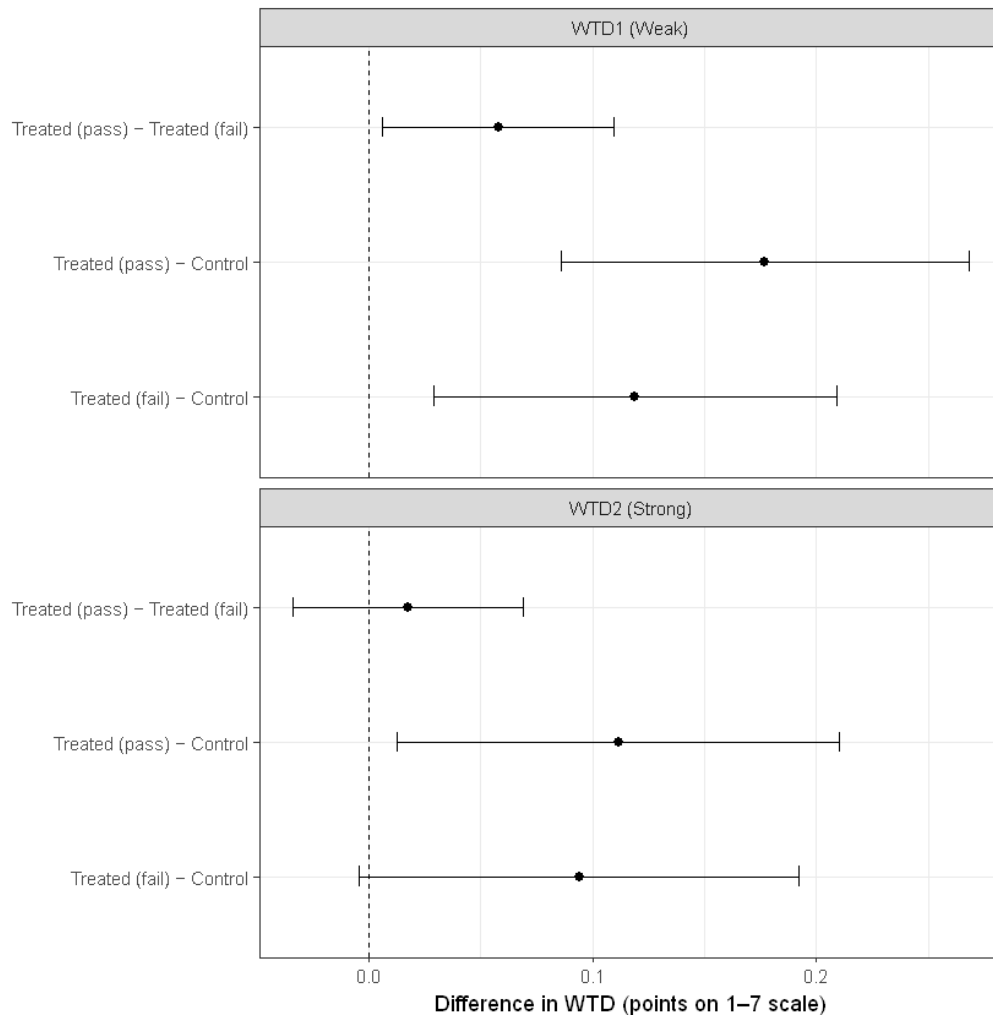
To summarize subgroup-specific effects, we report treatment-control contrasts within each political group. Inference is based on respondent-clustered robust standard errors to account for within-respondent dependence induced by the repeated-measures design. For WTD1, explanations increase willingness to delegate among right-leaning respondents ($= 0.313$, $SE = 0.060$, 95% CI $[0.194, 0.431]$, $p < 0.001$) and also among left-leaning respondents ($= 0.154$, $SE = 0.064$, 95% CI $[0.028, 0.279]$, $p =$

0.017). Effects for centrist respondents are imprecisely estimated (estimate = 0.155, SE= 0.111, 95% CI $[-0.064, 0.373]$), and estimates for the non-aligned/other group are close to zero ($= -0.057$, SE= 0.070, 95% CI $[-0.195, 0.081]$).

For WTD2, the treatment effect is concentrated among right-leaning respondents ($= 0.282$, SE= 0.067, 95% CI $[0.151, 0.413]$, $p < 0.001$), while effects in the Left, Center, and Non-aligned/Other groups are small and statistically indistinguishable from zero. Overall, these exploratory results suggest that explanations increase willingness to delegate primarily among respondents on the political right, with more limited evidence of an effect for weak delegation among left-leaning respondents.

Heterogeneity by attention

Figure 4.6: Attention to the explanation and willingness to delegate, pooled



A natural interpretation of the treatment is that it operates through comprehension of the procedural content. We therefore examine whether treatment effects are larger among treated respondents who

correctly answer two comprehension questions. Because the attention-check questions are only asked in the treated condition, control respondents cannot “pass” by construction. We therefore re-parameterize treatment into a three-level indicator: *Control*, *Treated (fail)*, and *Treated (pass)* (pass = 1 if both questions are answered correctly). We estimate pooled mixed-effects models with system fixed effects, the standard demographic controls, and respondent random intercepts, and compute cluster-robust standard errors clustered at the respondent level.

Figure 4.6 reports the implied contrasts for weak (WTD1) and strong (WTD2) delegation. For WTD1, explanations increase willingness to delegate among treated respondents who fail the attention check relative to the control group (Treated (fail) – Control: 0.119, SE = 0.046, $p = 0.010$). Importantly, respondents who pass the attention check exhibit an additional increase relative to treated failers (Treated (pass) – Treated (fail): 0.058, SE = 0.026, $p = 0.028$), yielding a larger overall effect for passers relative to controls (Treated (pass) – Control: 0.177, SE = 0.047, $p < 0.001$).

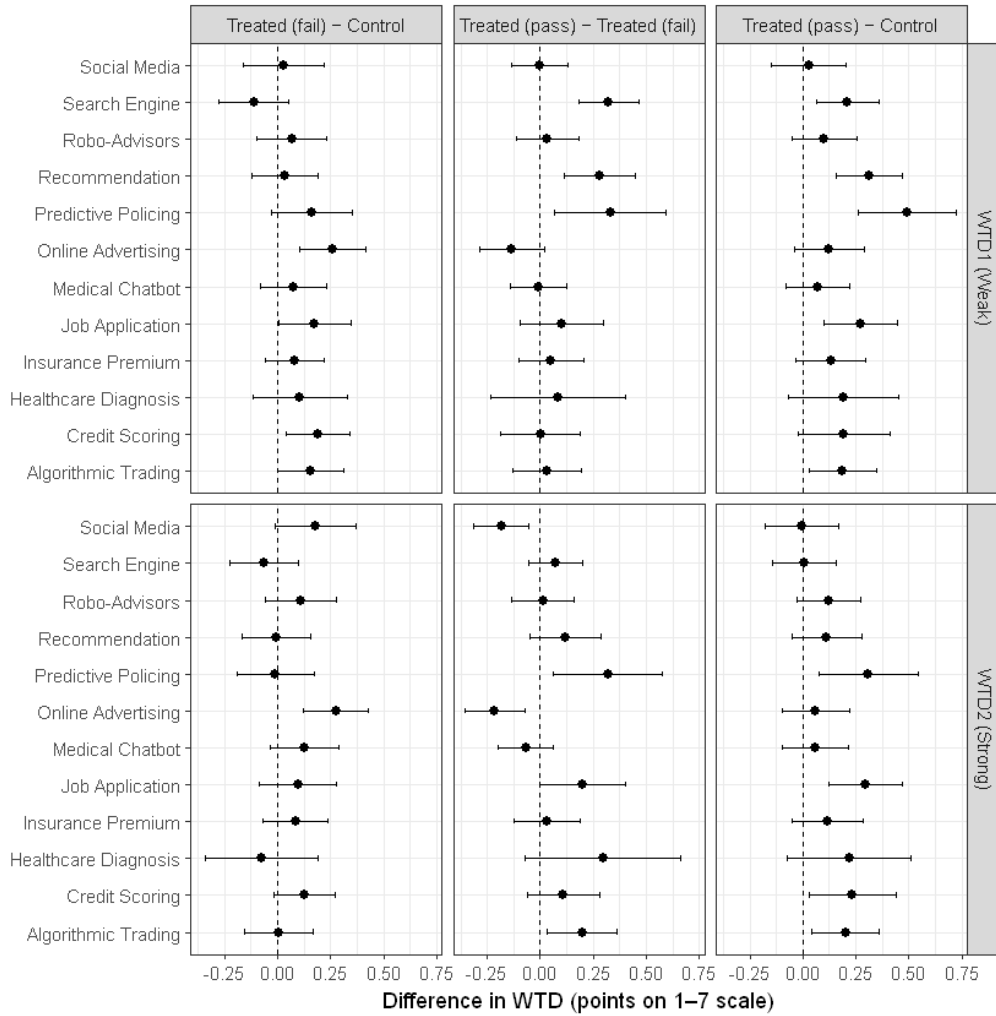
For WTD2, the treatment effect is weaker and less precisely estimated. Treated failers are only marginally higher than controls (Treated (fail) – Control: 0.094, SE = 0.050, $p = 0.061$). Passing the attention check does not generate an additional gain beyond treated failers (Treated (pass) – Treated (fail): 0.018, SE = 0.026, $p = 0.503$), although passers still differ positively from controls overall (Treated (pass) – Control: 0.112, SE = 0.051, $p = 0.027$).

These patterns suggest that attention to the explanatory content is associated with a larger increase in weak delegation, while differences between passers and failers are not detectable for strong delegation. As with the other exploratory analyses, these comparisons should be interpreted descriptively rather than as causal effects of “paying attention,” since attention-check performance is not randomly assigned. These comparisons quantify how the treatment effect varies with an observed post-treatment marker of comprehension, rather than the causal effect of comprehension itself.

We next replicate this analysis separately by system by interacting the three-level comprehension indicator (Control, Treated–fail, Treated–pass) allowing the 3-level indicator effects to vary by system, and we report three contrasts within each system: (i) Treated (fail) - Control, (ii) Treated (pass) - Treated (fail), and (iii) Treated (pass) - Control. Figure 4.7 displays point estimates and 95% confidence intervals using respondent-clustered standard errors. Because these comparisons are exploratory and numerous (12 systems $\times$ 3 contrasts $\times$ 2 outcomes), we interpret patterns qualitatively and treat conventional p -values as descriptive.

For weak delegation (WTD1), the treated–control gaps are positive in many systems, but the magnitude varies substantially. The largest “Treated (pass) - Control” effects occur for Predictive Policing (+0.493) and Recommendation (+0.314), and are also sizable for Job Applications (+0.273). Several additional systems show more modest positive effects (e.g., Algorithmic Trading and Credit Scoring).

Figure 4.7: Attention to the explanation and willingness to delegate, by system



Differences between treated passers and treated failers are only statistically distinguishable from zero in a small subset of systems. In particular, passers report significantly higher willingness to delegate than failers for Predictive Policing (+0.332) and Recommendation (+0.280), suggesting that better comprehension is associated with greater acceptance in these domains. By contrast, Online Advertising exhibits the opposite pattern: while Treated (fail) exceeds Control (+0.259), Treated (pass) is not reliably above Control (+0.125) because the “pass-fail” contrast is negative (-0.134).

For strong delegation (WTD2), effects are more selective and generally less precisely estimated. The clearest positive “Treated (pass) - Control” contrasts appear for Predictive Policing (+0.309) and Job Applications (+0.296). As with WTD1, the “pass-fail” contrast is positive for Predictive Policing (+0.319), while it is negative for Online Advertising (-0.214) and Social Media (-0.183). In both of these latter systems, treated failers are above (or marginally above) controls, whereas passers are approximately indistinguishable from controls, consistent with the idea that a more complete understanding may attenuate

willingness to fully delegate in advertising- and social-media contexts.

Overall, the system-level results suggest that the relationship between comprehension and delegation is not uniform across domains: passing the attention check is associated with substantially higher willingness to delegate for some systems (most notably predictive policing and recommendation), but not for others, where better comprehension may instead coincide with greater caution (online advertising and social media).

4.4.5 Summary of results

Across the 12 algorithmic domains, providing respondents with a brief explanation of how each system works produces clear and consistent shifts in subjective beliefs, more modest improvements in evaluative attitudes, and small but detectable increases in willingness to delegate.

Descriptive patterns. In the baseline condition, beliefs and delegation vary substantially across systems. Everyday digital technologies (e.g., search engines and social media) are rated highest on familiarity, exposure, and use, whereas institutionally distant applications (e.g., predictive policing) score lowest. Attitudes are mildly favorable on average, especially for fairness, trust, and usefulness, while transparency remains the least favorable dimension.

Reduced-form treatment effects. Pooling across systems, explanations significantly increase all belief items—knowledge, familiarity, perceived exposure, and self-reported use—by roughly 0.15 to 0.25 points on the 1–7 scale. Attitudinal effects are positive but smaller and concentrated in fairness, trust, usefulness, and (to a lesser extent) accountability; transparency concerns remain essentially unchanged, and privacy responses are mixed across items. Delegation rises modestly: weak delegation increases more reliably than strong delegation.

Mechanism-consistent decomposition. A recursive three-equation decomposition indicates that the total treatment effect on delegation is primarily consistent with pathways operating through attitudes. The direct treatment component is close to zero, while most of the increase in both weak and strong delegation is accounted for by (i) an attitudes-only pathway and (ii) a smaller sequential pathway from beliefs to attitudes to delegation. These conclusions are robust to constructing beliefs and attitudes indices either as sums or via PCA.

Exploratory heterogeneity. Item-level estimates confirm that belief gains are largest for perceived exposure and self-reported use, and that evaluative improvements are concentrated in fairness, trust, and usefulness. Using a control-group proxy for baseline system familiarity, conditional effects are broadly positive at both low and high familiarity, with limited evidence of systematic moderation. Political heterogeneity is more pronounced for strong delegation: effects are concentrated among right-leaning respondents, whereas weak delegation increases among right-leaning respondents and more modestly

among left-leaning respondents. Finally, using an attention-check measure asked only in the treated condition, delegation increases even among treated respondents who fail the check, while passing the check is associated with an additional gain for weak delegation (but not for strong delegation) in the pooled analysis; system-specific patterns are heterogeneous, with larger pass-fail gaps in some domains (e.g., predictive policing and recommendation) and negative gaps in others (e.g., online advertising and social media). Because these heterogeneity analyses rely on post-treatment subgrouping and many comparisons, they are interpreted descriptively.

Taken together, the results show that brief explanations reliably increase perceived understanding and improve several evaluative attitudes, and that these attitudinal shifts account for most of the modest downstream increase in willingness to delegate.

4.5 Conclusion

Algorithmic decision-making systems increasingly mediate access to jobs, credit, insurance, public services, and information. In Europe, this shift is occurring alongside an expanding legal architecture that explicitly ties the legitimacy of algorithmic decision-making to information disclosure and transparency requirements.⁴ Yet there remains a basic, economically salient empirical question: when people receive the kind of “light-touch” explanation that regulators can realistically mandate at scale, do they update beliefs and attitudes in ways that plausibly translate into adoption and delegation?

We answer this question with a large survey experiment in three major EU countries (France, Germany, and Italy; $N = 3007$; Table 4.A.4). The design deliberately targets a policy-relevant margin. All respondents receive a neutral description of what a system is used for; the treatment adds a short, non-technical explanation of how it works. Because each respondent evaluates multiple algorithmic domains, we can compare reactions across contexts while holding fixed individual traits, and we can distinguish “weak” delegation (using the system as decision support) from “strong” delegation (ceding decision rights).

Three results stand out. First, explanations generate large and robust shifts in perceived familiarity and experience. Relative to the baseline, treated respondents report higher knowledge, familiarity, exposure, and self-reported use (Table 4.2), with average differences on the order of 0.2–0.5 points on a 7-point scale. This is consistent with a central premise from information economics: even simple disclosures can relax informational frictions by helping people form a more concrete mental model of an otherwise opaque technology.

Second, explanations improve a broad set of evaluative attitudes that are central to both policy debates and adoption decisions. Pooling across domains, the treatment increases perceived fairness, accountabil-

⁴See, among others, the EU AI Act (Regulation (EU) 2024/1689) and the Digital Services Act (Regulation (EU) 2022/2065).

ity, trust, and usefulness, and it modestly shifts privacy-related responses (Table 4.3). Transparency itself appears harder to move: one transparency item is essentially unaffected and the other shifts slightly in the opposite direction, suggesting that brief descriptions of “how it works” may raise standards for what respondents consider genuinely transparent. Put differently, disclosure can improve perceived legitimacy along several margins while still leaving a residual “opacity” concern: a pattern that is highly relevant for policy designs that equate disclosure with transparency.

Third, explanations increase willingness to delegate, but the effect is modest and concentrated in weak delegation. Average willingness to use the system as assistance rises by about 0.11 points, while willingness to cede decision rights rises by roughly 0.06 points (Table 4.4). This gap matters economically: it implies that the behavioral margin most responsive to realistic disclosures is complementarity between humans and algorithms (support), rather than full substitution (automation). In a European context where institutional legitimacy and accountability constraints are salient, this distinction helps reconcile why transparency mandates may increase acceptance without producing blanket enthusiasm for “turning decisions over” to machines.

Beyond reduced-form effects, our mechanism-consistent decomposition reinforces an economic interpretation of these patterns. The direct effect of receiving an explanation on delegation is close to zero, while the total effect is largely mediated through updated beliefs and, especially, through updated evaluative attitudes (Table 4.5). This structure is informative for both theory and policy: disclosures appear to work less by mechanically “persuading” respondents and more by shifting the perceived trade-offs that underpin adoption decisions (e.g., usefulness versus privacy concern, or fairness and accountability versus opacity).

The cross-domain nature of the study also highlights a key takeaway for European governance: public acceptance is not a single parameter. Even under a standardized disclosure, the responsiveness of beliefs, attitudes, and delegation varies across application areas, consistent with the idea that perceived stakes, contestability, and prior salience shape how information is processed. This heterogeneity cautions against one-size-fits-all disclosure templates. It also suggests a practical path forward for regulators and firms: if the goal is legitimate uptake, disclosures should be targeted to the dimensions that are most binding in a given domain (e.g., data provenance and privacy in some settings; accountability in others), rather than relying on generic “how it works” language.

Taken together, the paper offers an empirically grounded, Europe-relevant perspective on the economics of algorithmic legitimacy. Our central contribution is to evaluate a disclosure intervention that is (i) lightweight enough to be scalable and legally realistic, (ii) comparable across a wide range of economically important ADM domains, and (iii) measured on outcomes that map directly to adoption and delegation decisions. This combination is what makes the evidence particularly useful for European

scientific audiences and policy debates: it bridges legal requirements (disclosure and transparency) and economic mechanisms (belief updating, attitude formation, and delegation choices), and it does so in the population and institutional context where many of these rules are being implemented.

Several limitations point to clear directions for future research. First, our outcomes are stated preferences in a survey setting; field evidence on revealed adoption and realized welfare remains essential. Second, we study one standardized form of explanation; richer disclosures (or interactive explanations) may shift the transparency margin differently, but they may also impose cognitive costs and create new forms of persuasion or overreliance. Third, disclosures are only one component of governance: future work should examine how explanation mandates interact with auditing, contestability rights, and liability regimes, and whether these bundles differentially affect trust, compliance, and inequality. These extensions are not only policy-relevant; they are also fundamental to understanding when algorithmic systems function as productivity-enhancing complements and when they become contested substitutes.

In the near term, our results support a tempered conclusion: realistic explanations are not a panacea, but they are not empty formalities either. They reliably shift beliefs, improve several key evaluative attitudes, and modestly increase willingness to rely on algorithms mainly as decision support. For European policymakers concerned with the legitimacy of algorithmic decision-making, this suggests that well-designed disclosure requirements can be a low-cost complement to stronger institutional safeguards, helping align adoption with informed acceptance rather than blind trust or blanket rejection.

4.6 Appendix

4.7 Survey design appendices

4.7.1 Items

Table 4.A.1: Initial Beliefs Items

Construct	No.	Measurement item	Ref.
Self-Reported Knowledge	K1	I have heard or read about the use of these types of systems that make automated decisions or suggestions. ^a	Aysolmaz et al. (2023)
	K2	I am familiar with these types of systems that make automated decisions or suggestions. ^a	
Perceived exposure	EXP1	I believe I am exposed to these types of systems. ^b	
Self-Reported Use	U1	I believe I use these types of systems. ^b	

^a Response options: Strongly disagree; Disagree; Somewhat disagree; Neutral; Somewhat agree; Agree; Strongly agree.

^b Response options: Never; Once a decade; Once a year; Once a month; Weekly; Daily; Multiple times in the day.

Table 4.A.2: Attitude and Willingness to Delegate Items

Construct	No.	Measurement item	Ref.
Transparency Concerns	TC1	I am concerned that it would not be exactly clear how decisions or suggestions are produced by this system. ^a	Sutanto et al. (2013); W. Wang and Benbasat (2016)
	TC2	It would bother me if this system does not provide a sufficient explanation about how certain decisions are taken or suggestions are given to me. ^a	–
Perceived Fairness	PF1	I believe that this system in general would treat me fairly when making decisions and suggestions. ^a	M. K. Lee (2018)
	PF2	I think that this system would be objective in making decisions or suggestions for me. ^a	–
Perceived Accountability	PA1	If I have any problem with the decisions or suggestions of this system, I believe that the organization providing the system would take necessary measures. ^a	Kacianka and Pretschner (2021); Wieringa (2020)
	PA2	I think that if this system causes harm and nuisance for me then this would be compensated by the organization providing the system. ^a	–
Perceived Usefulness	PU1	I think that I am better off with the decisions and suggestions made by the system. ^a	Gefen et al. (2003)
	PU2	I feel that using this system would make my life more efficient by taking decisions or providing suggestions for me. ^a	–
Privacy Concern	PC1	I am concerned about the use of my personal information by this system. ^a	Bansal et al. (2015)
	PC2	I believe that the privacy of my personal data would be assured while using this system. ^a	–
Perceived Trust	Tr1	I trust that this system would provide high-quality decisions about me. ^a	Jian et al. (2000); M. K. Lee (2018)
	Tr2	I believe that this system would not intentionally harm me. ^a	–
Willingness to delegate	WTD1	I would be willing to let this system help/assist me. ^a	Gefen et al. (2003)
	WTD2	I would be willing to delegate my decision rights to this system. / I am willing to let this system make decisions for me. ^a	

^a Response options: Strongly disagree; Disagree; Somewhat disagree; Neutral; Somewhat agree; Agree; Strongly agree.

4.7.2 Definitions and Explanations

Table 4.A.3: System types with definitions and explanations

Type	Definition	Explanation
Credit Scoring Algorithms	A credit scoring algorithm is a tool banks and lenders use to quickly evaluate how likely someone is to repay a loan. It assigns a numeric score indicating creditworthiness based on the person’s financial history.	A credit scoring algorithm evaluates an individual’s financial history, including debt repayment, credit usage, loan history, and length of credit accounts. Each factor is assigned a numerical value based on its importance in predicting repayment behavior. These values are combined, using weighted calculations, to generate a single numerical score. The algorithm applies statistical methods to ensure the score accurately reflects financial reliability. Finally, this numeric score summarizes the individual’s overall creditworthiness.
Online Advertising Algorithms	An online advertising algorithm is a digital system that decides which ads users see on websites or social media platforms. It selects ads based on user interests and behaviors to make advertisements relevant.	An online advertising algorithm tracks user behaviors, such as website visits, clicks, searches, and demographic details. It builds detailed user profiles by analyzing patterns from this collected data. Based on these profiles, the algorithm selects ads that closely align with user interests or predicted preferences. It determines the optimal placement and timing for displaying each advertisement across websites or applications. The algorithm continuously learns and adjusts its ad selections by monitoring user responses and interactions.

Continued on next page

Type	Definition	Explanation
Algorithmic Trading	Algorithmic trading is an automated way of buying and selling financial assets like stocks using computer programs. These programs execute trades quickly based on predefined rules to maximize profit or manage risk.	Algorithmic trading uses automated computer programs that monitor real-time market data, such as price movements, volume, and trends. These programs operate based on predefined rules and trading strategies programmed into the system. When market conditions match these programmed criteria, the algorithms automatically execute buy or sell orders. Trades are performed rapidly, often within milliseconds, without human input. The algorithm continually analyzes market data to adjust or close positions according to its instructions.
Robo-Advisors	A robo-advisor is an automated online service that manages investments and financial planning using computer algorithms. It provides affordable, personalized investment advice without human intervention.	A robo-advisor gathers basic information about a user's financial situation and investment goals, and then applies automated algorithms to analyze this data. It uses these algorithms to select and manage investments, typically assembling them into diversified portfolios. The system continuously monitors market conditions and automatically adjusts the investments as necessary. It regularly provides updates and reports to inform users of their investment progress.

Continued on next page

Type	Definition	Explanation
Insurance Premium Calculation Algorithms	An insurance premium calculation algorithm determines how much a person should pay for insurance coverage. It sets premiums based on factors that predict the likelihood and cost of potential claims.	An insurance premium calculation algorithm analyzes data such as age, health history, driving behavior, or property characteristics. It assigns numerical risk values to these factors based on their likelihood of causing future claims. Each factor is weighted according to its significance in influencing potential claim costs. The algorithm combines these weighted factors to compute a total risk score. This score directly translates into the premium amount charged to the policyholder.
Healthcare Diagnosis Algorithms for Experts	Healthcare diagnosis algorithms are computational tools designed to support medical professionals by systematically interpreting clinical data. They aid in identifying diseases or conditions by analyzing symptoms, test results, and patient history.	Healthcare diagnosis algorithms process patient data, including symptoms, medical history, lab results, and imaging. They compare this data against extensive medical databases and clinical guidelines. Algorithms identify patterns by matching patient information with known diagnostic criteria. Through statistical or machine learning methods, they calculate the likelihood of different medical conditions. The algorithm then provides clinicians with potential diagnoses ranked by probability.

Continued on next page

Type	Definition	Explanation
Medical Assistant or Medical Chatbot	A medical assistant or medical chatbot is an automated digital tool designed to interact with users about health-related questions or concerns. It provides instant guidance, symptom assessment, or basic medical information through conversational interactions.	A medical assistant or chatbot receives user input in the form of text or voice-based questions. It uses algorithms to interpret and understand these inputs through natural language processing. The chatbot matches the interpreted query to relevant medical information stored in its database. It generates responses based on predefined guidelines or machine learning models trained on clinical data. Finally, it communicates its response back to the user in a conversational format.
Job Application Screening Algorithms	A job application screening algorithm is an automated tool employers use to quickly evaluate job applicants. It filters and ranks candidates based on their resumes, experience, skills, and qualifications.	A job application screening algorithm scans resumes and applications submitted by candidates. It identifies and extracts keywords, qualifications, experience, education, and skills mentioned within these documents. The algorithm compares extracted information against predefined criteria from the job description. Applicants are ranked or scored based on how closely their profiles match the required criteria. Finally, it generates a shortlist of candidates who best align with the job requirements.

Continued on next page

Type	Definition	Explanation
Predictive Policing Algorithms	A predictive policing algorithm is a computational tool used by law enforcement to anticipate where or when crimes might occur. It analyzes historical crime data to guide policing efforts proactively.	A predictive policing algorithm analyzes historical crime data, such as locations, times, types, and frequency of past incidents. It identifies patterns or trends within this data using statistical or machine-learning techniques. The algorithm then calculates probabilities to estimate where and when crimes might occur next. It generates maps or reports highlighting areas at higher risk of future crimes. These predictions guide police deployments and patrol strategies.
Recommendation Algorithms	A recommendation algorithm is an automated system that suggests items, content, or services to users based on their preferences or behaviors. It provides personalized recommendations, such as movies, products, or articles, to enhance user experience.	A recommendation algorithm tracks user interactions, such as clicks, views, ratings, or purchases. It analyzes this data to identify patterns and similarities among users or items. The algorithm applies statistical or machine learning methods to predict user preferences. Based on these predictions, it selects and ranks items that closely match individual user profiles. Finally, the algorithm presents personalized suggestions directly to the user interface.
Search Engine Algorithms	A search engine algorithm is a computerized method that finds and ranks relevant web pages in response to a user's query. It determines which results best match the search terms entered.	A search engine algorithm scans billions of web pages, collecting and indexing content such as text, images, and links. It analyzes user-entered keywords or phrases to determine their intent. The algorithm then compares these keywords against its indexed pages to find relevant matches. It ranks results based on factors like keyword relevance, page quality, popularity, and user engagement. Finally, the algorithm displays a list of web pages organized by their ranking.

Continued on next page

Type	Definition	Explanation
Social Media Algorithms	A social media algorithm is an automated system that selects and organizes content displayed on users' feeds. It determines which posts users see first based on their interests and behaviors.	A social media algorithm monitors user interactions such as likes, shares, comments, and views. It analyzes this data to build personalized profiles reflecting individual interests and behaviors. The algorithm ranks content by assessing its relevance to these user profiles. It prioritizes and arranges posts, stories, or videos accordingly within user feeds. The content displayed continuously updates based on ongoing user interactions.

4.8 Additional Results

4.8.1 Additional Figures

Beliefs Items

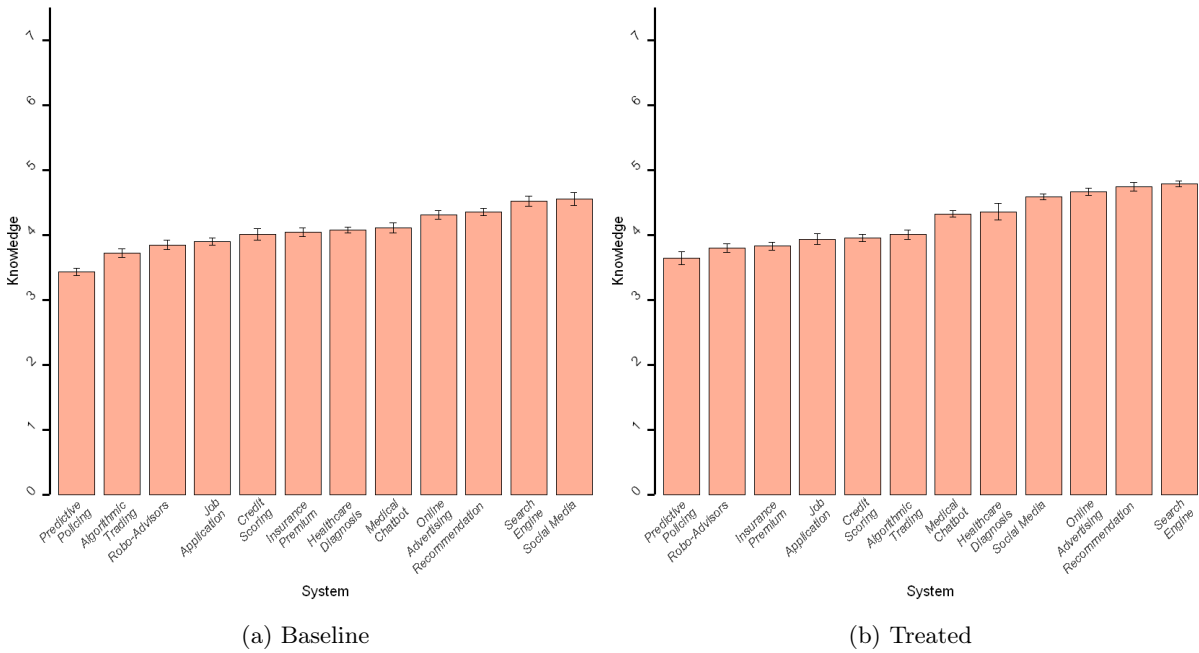


Figure 4.A.1: Self-Reported Knowledge

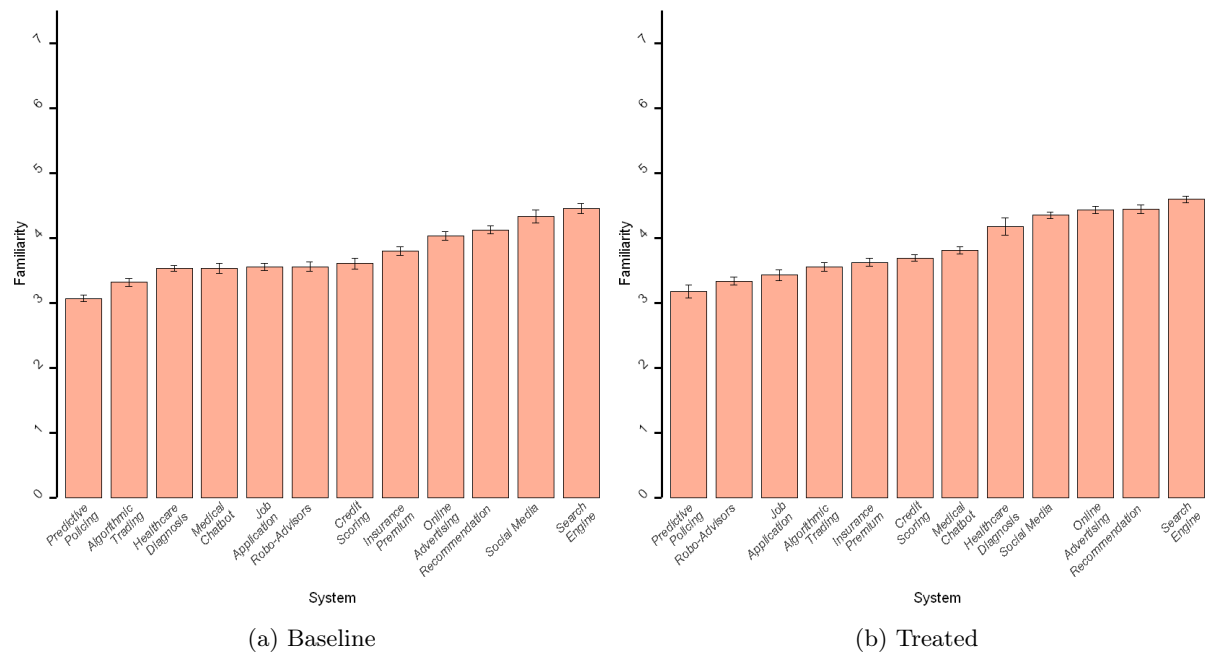


Figure 4.A.2: Self-Reported Familiarity

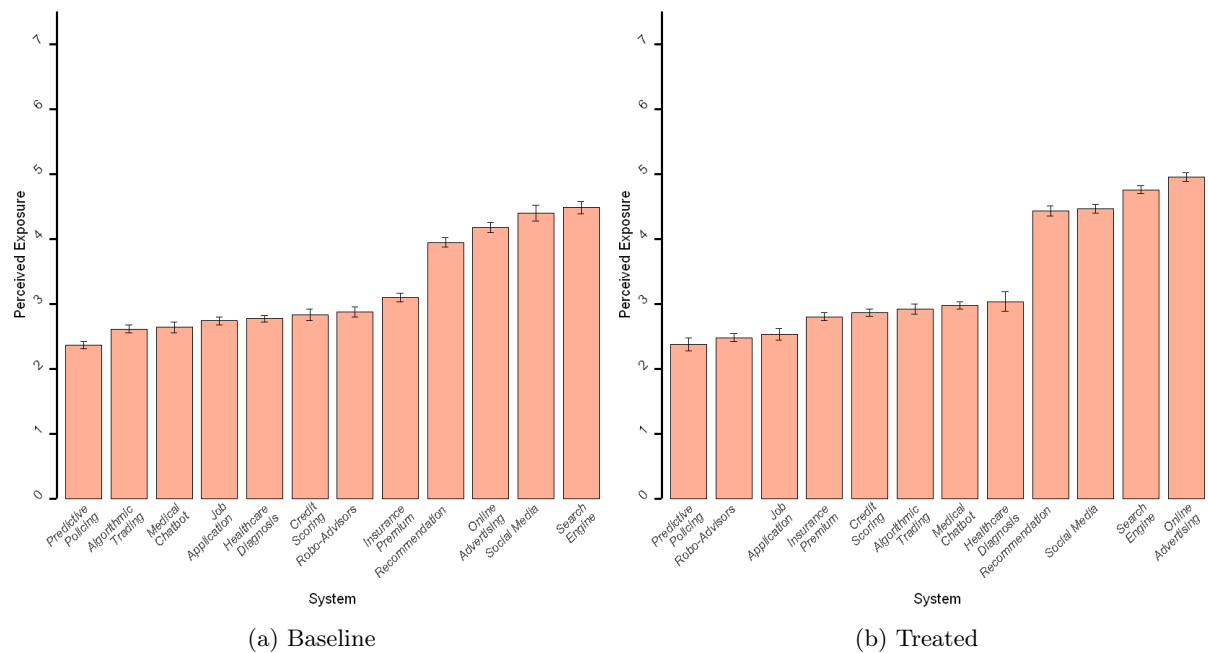


Figure 4.A.3: Perceived Exposure

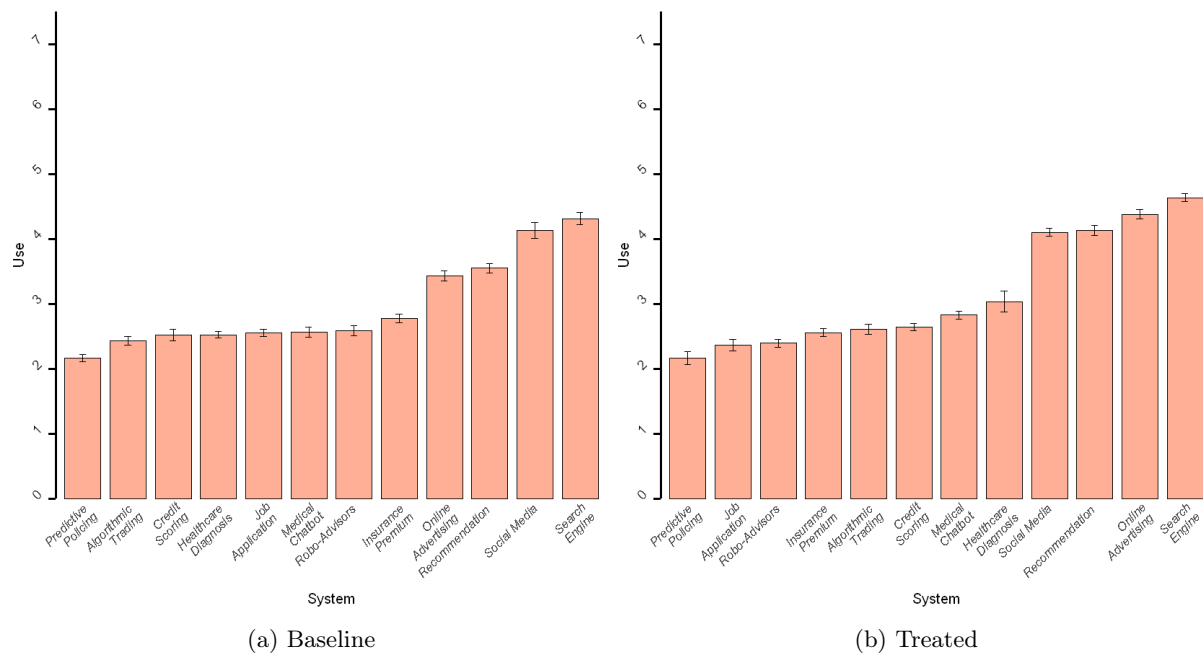


Figure 4.A.4: Self-Reported Use

Attitudes Items

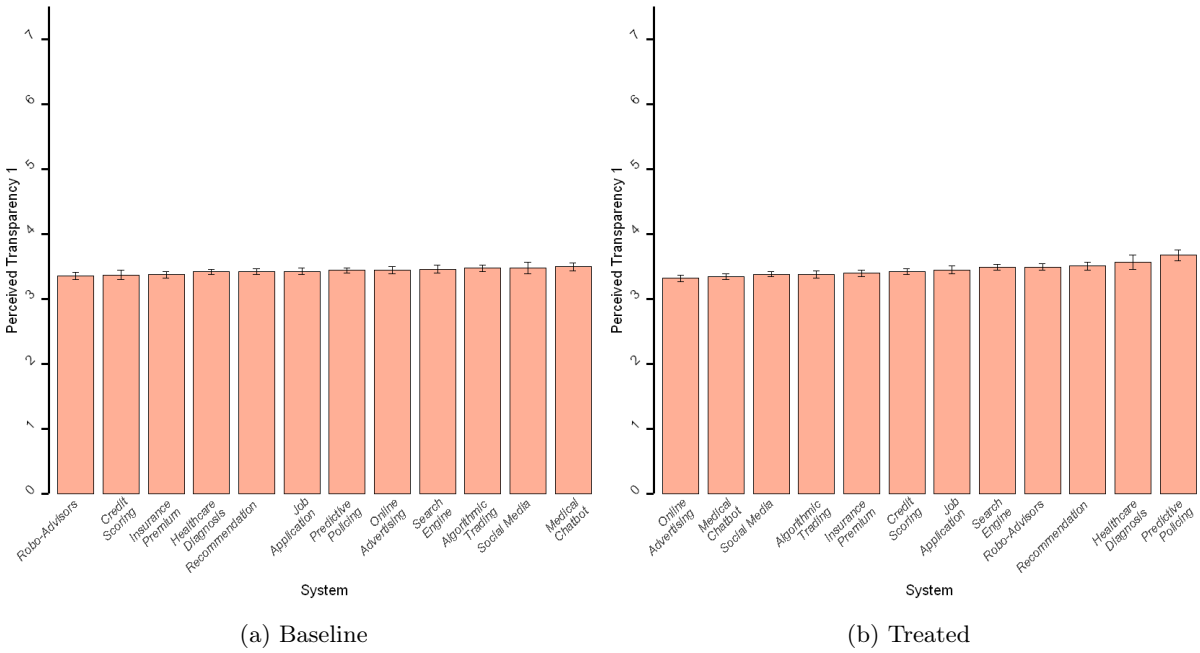


Figure 4.A.5: Transparency Concerns 1

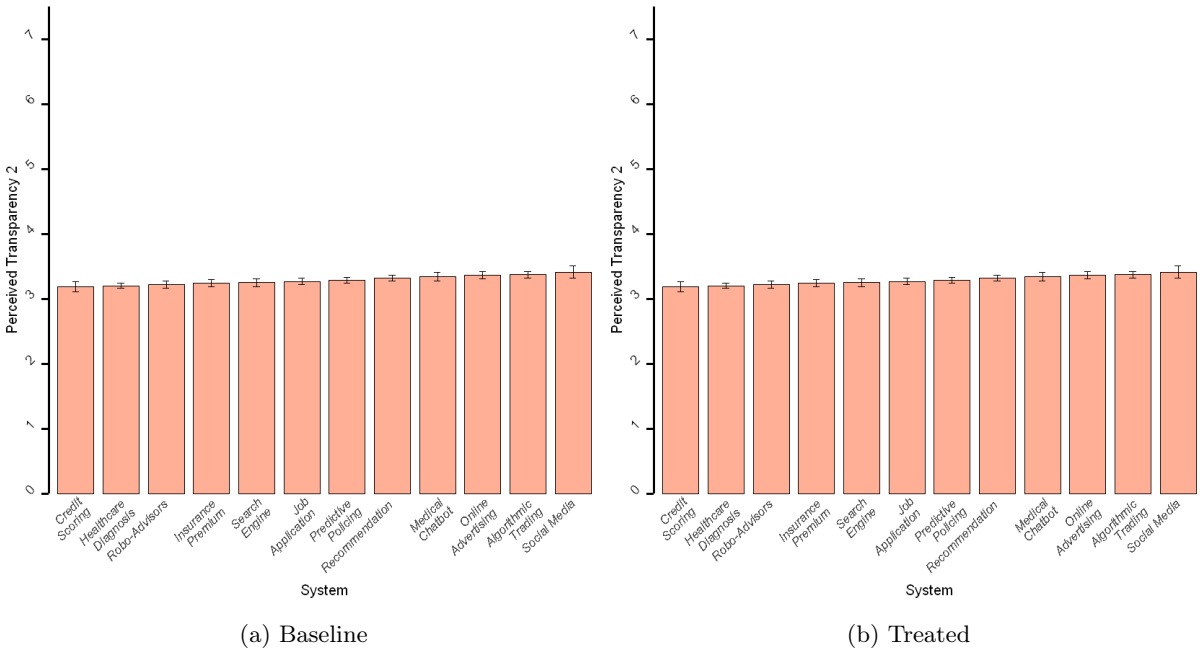


Figure 4.A.6: Transparency Concerns 2

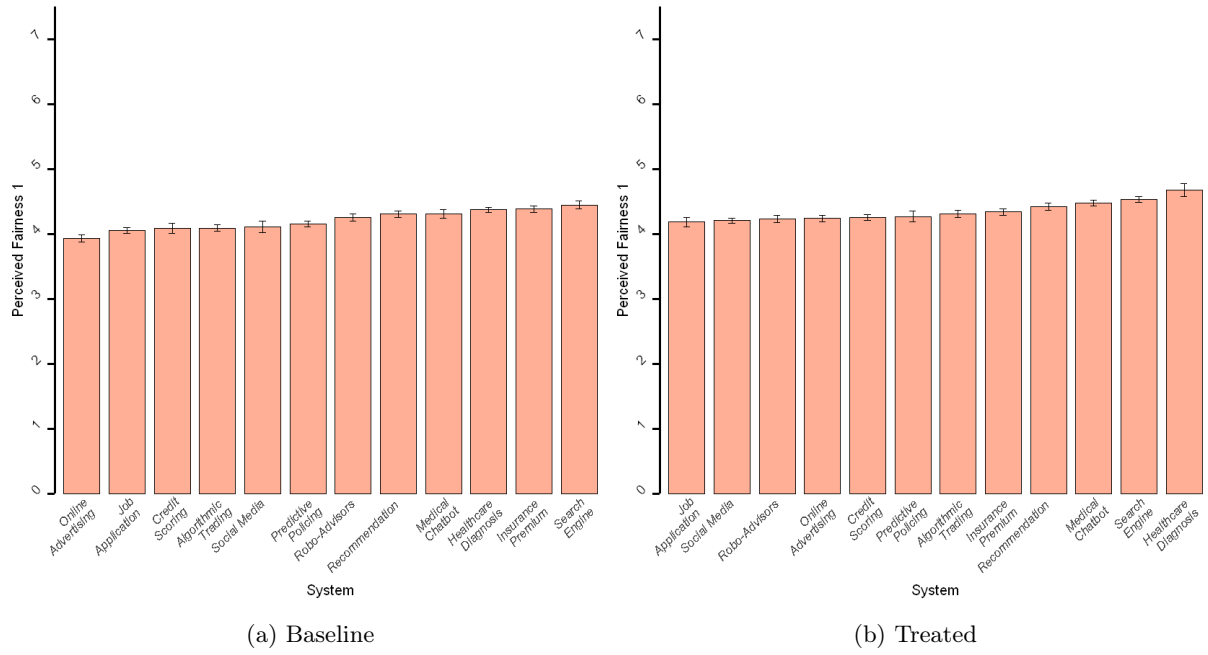


Figure 4.A.7: Perceived Fairness 1

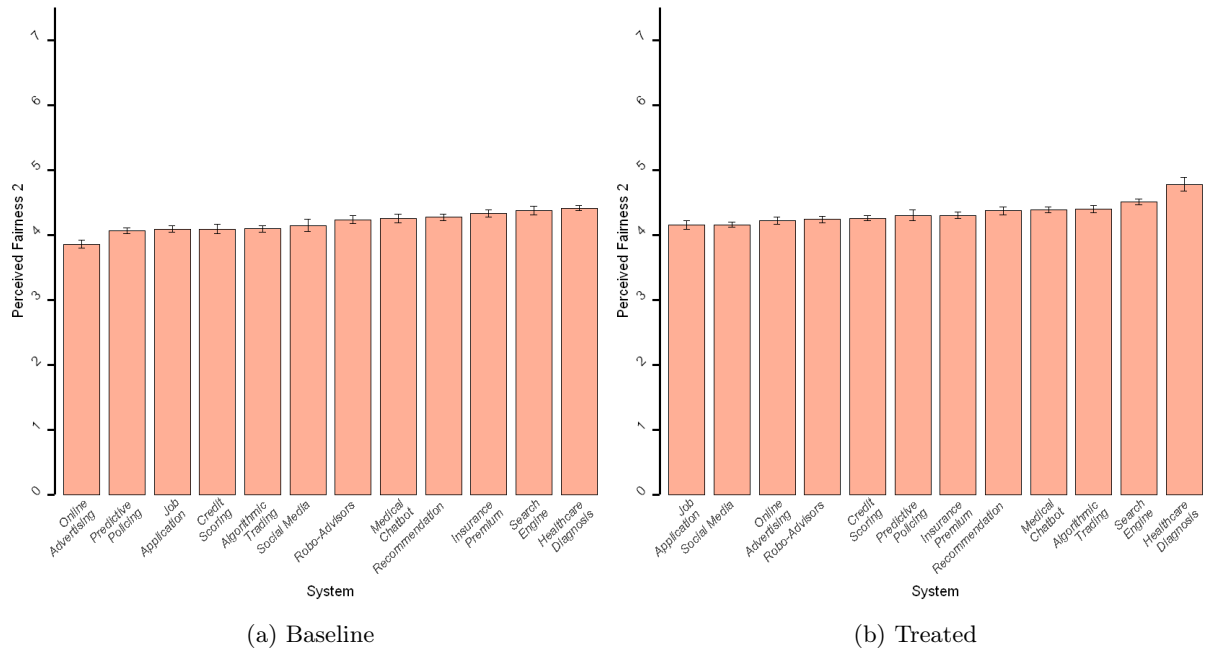


Figure 4.A.8: Perceived Fairness 2

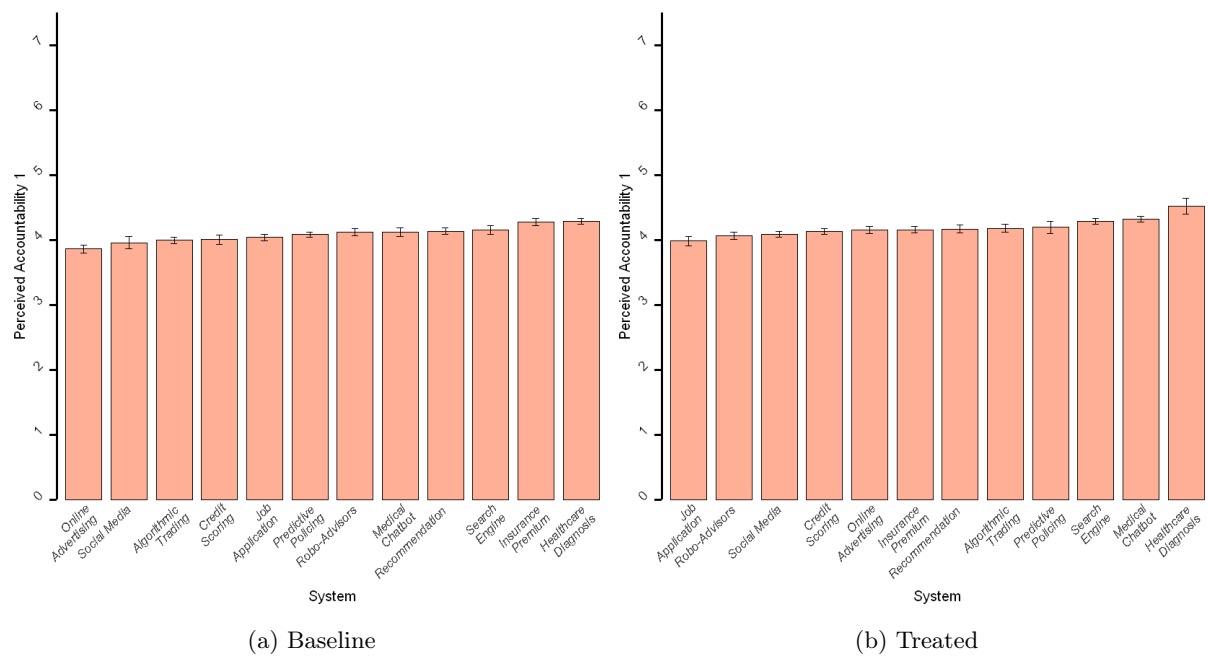


Figure 4.A.9: Perceived Accountability 1

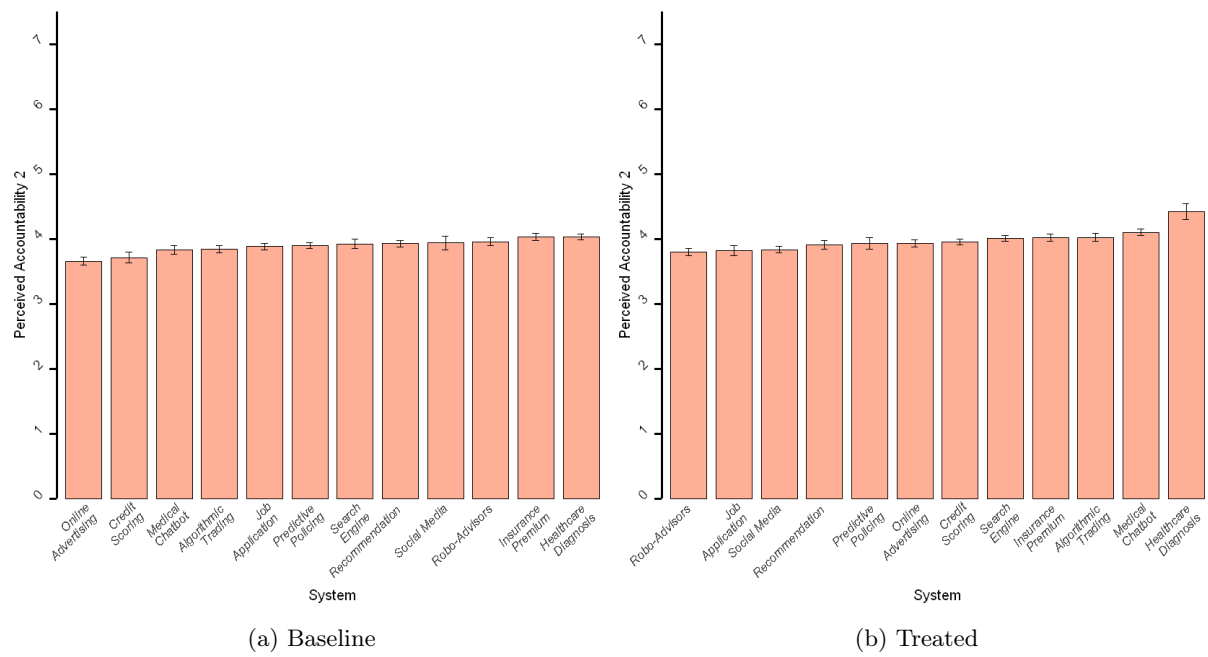


Figure 4.A.10: Perceived Accountability 2

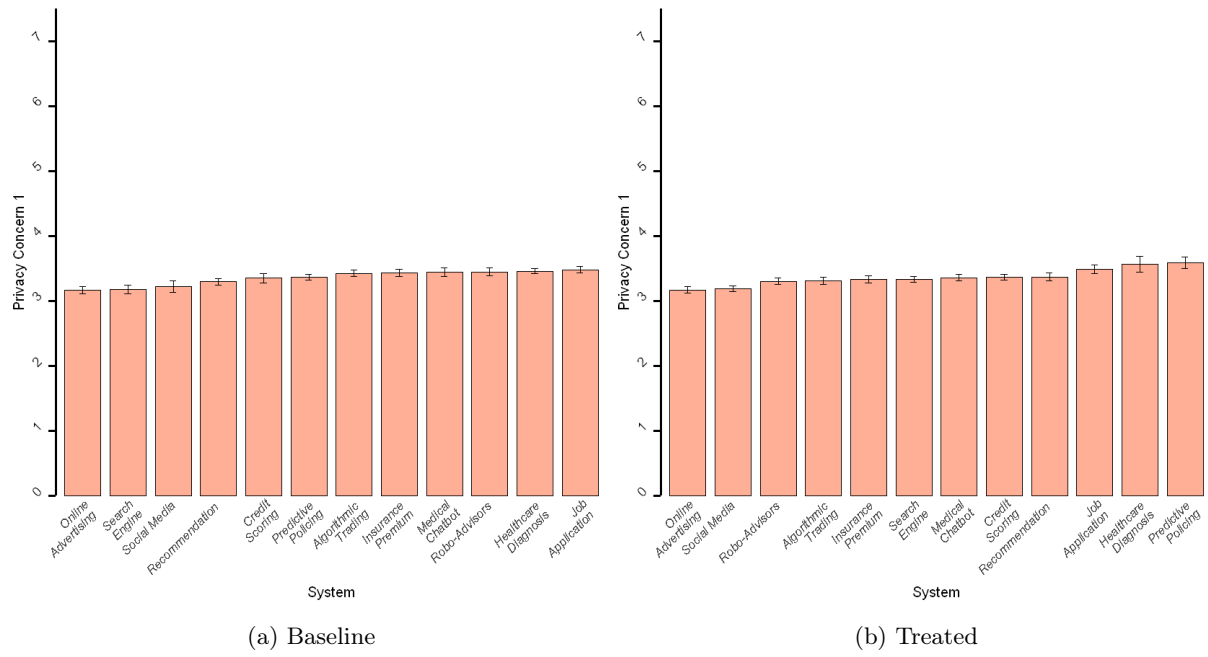


Figure 4.A.11: Privacy Concern 1

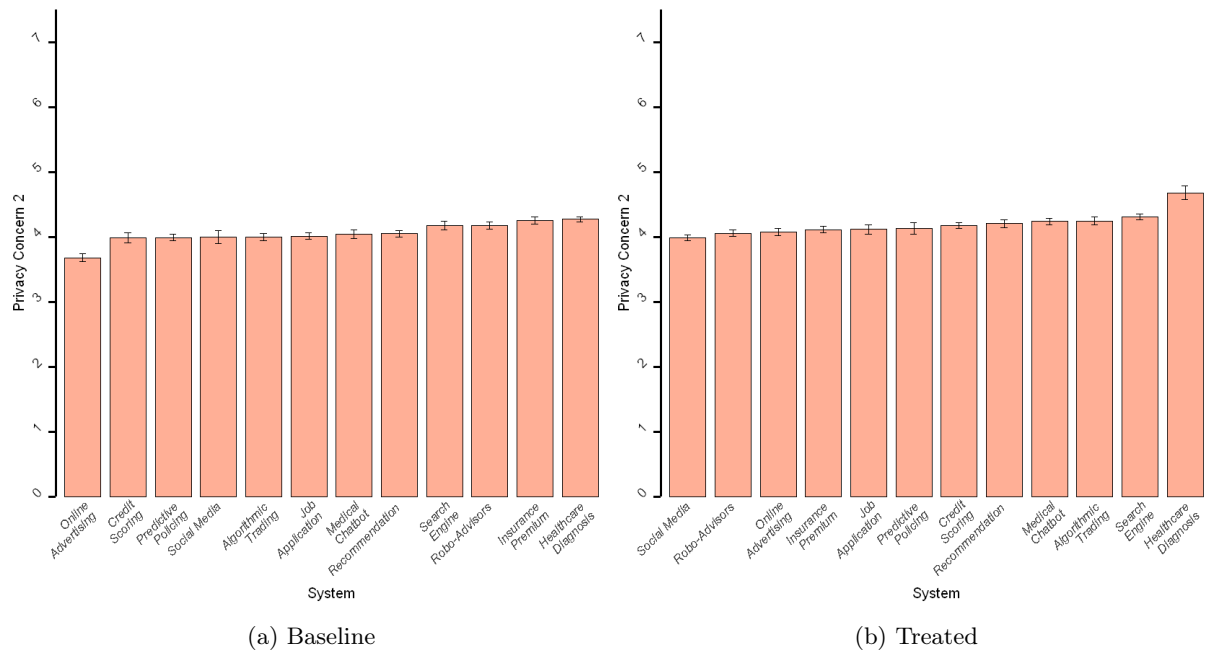


Figure 4.A.12: Privacy Concern 2

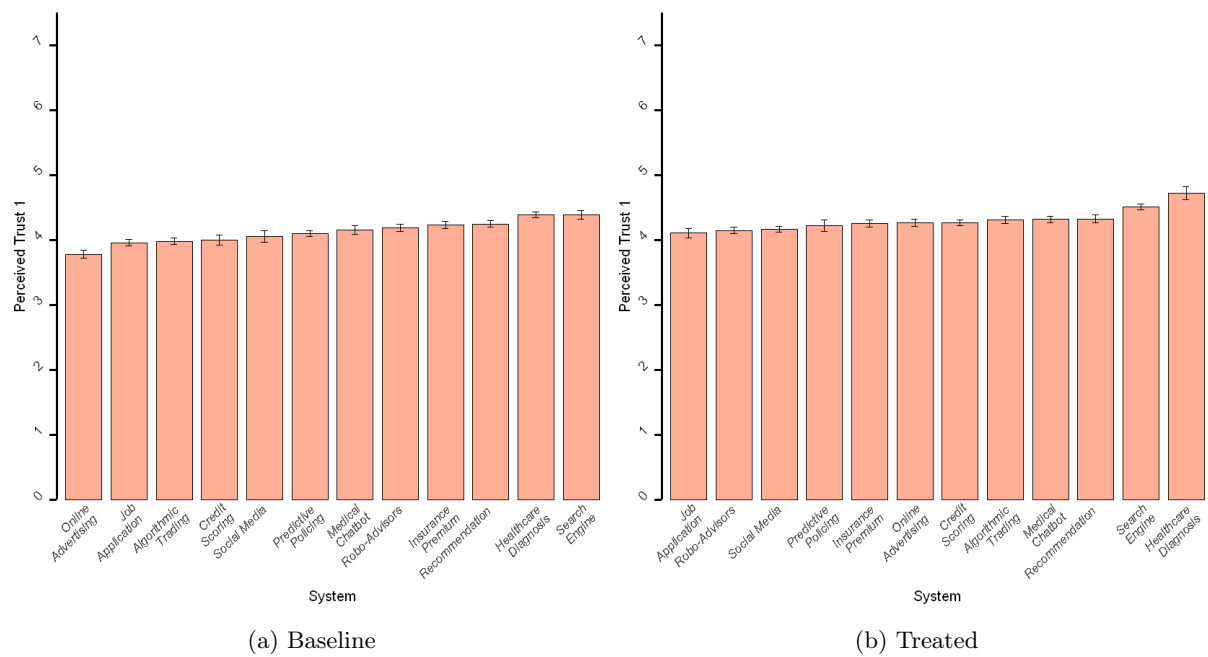


Figure 4.A.13: Perceived Trust 1

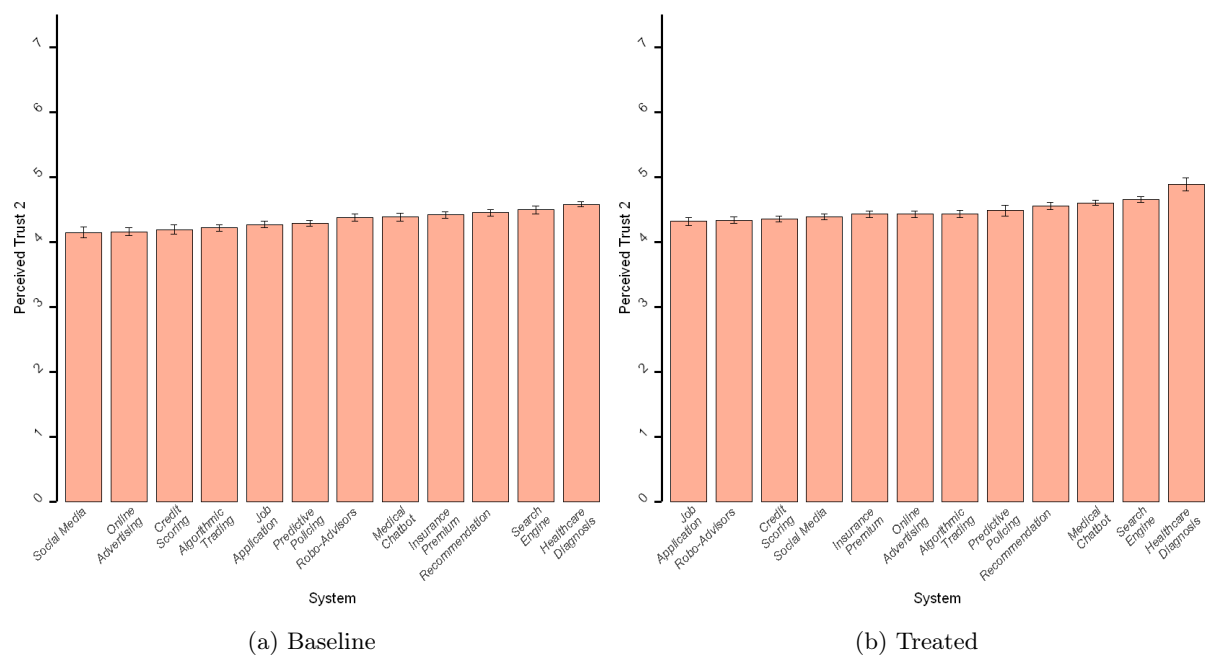


Figure 4.A.14: Perceived Trust 2

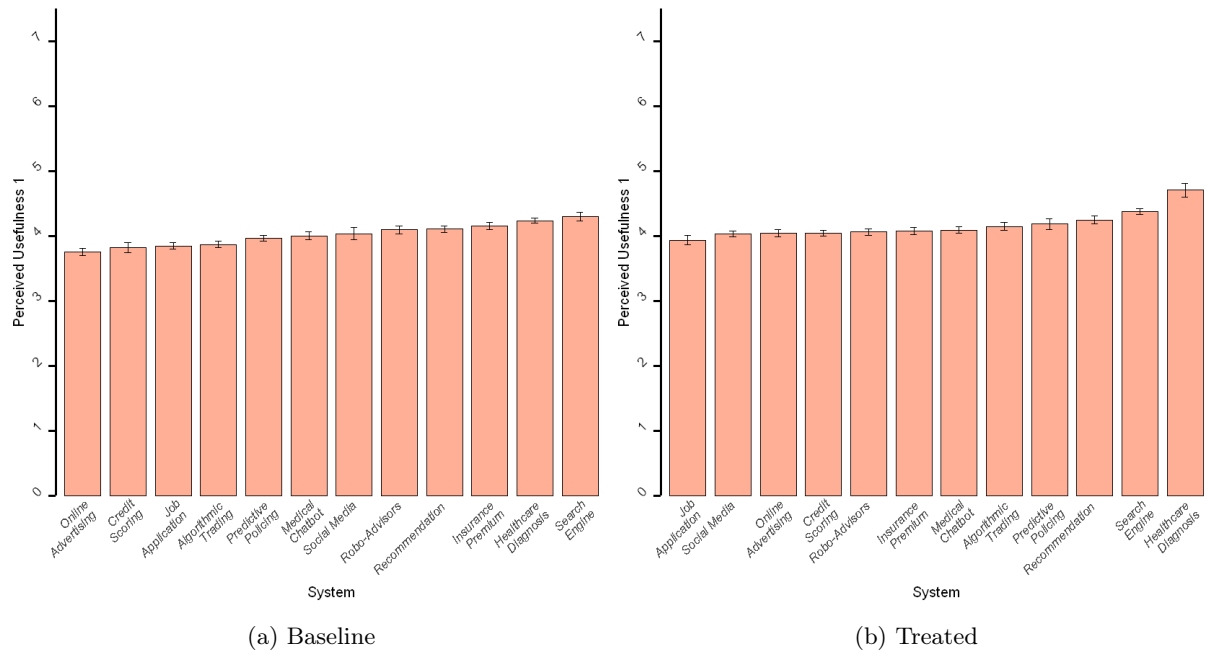


Figure 4.A.15: Perceived Usefulness 1

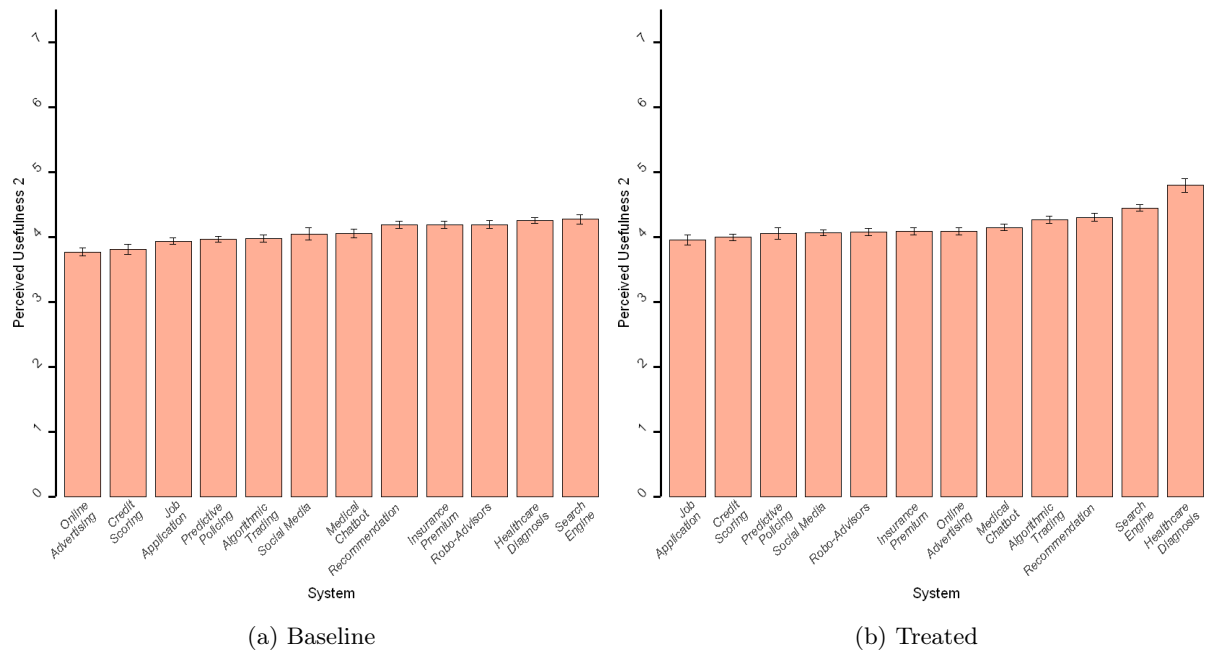


Figure 4.A.16: Perceived Usefulness 2

Willingness to Delegate Items

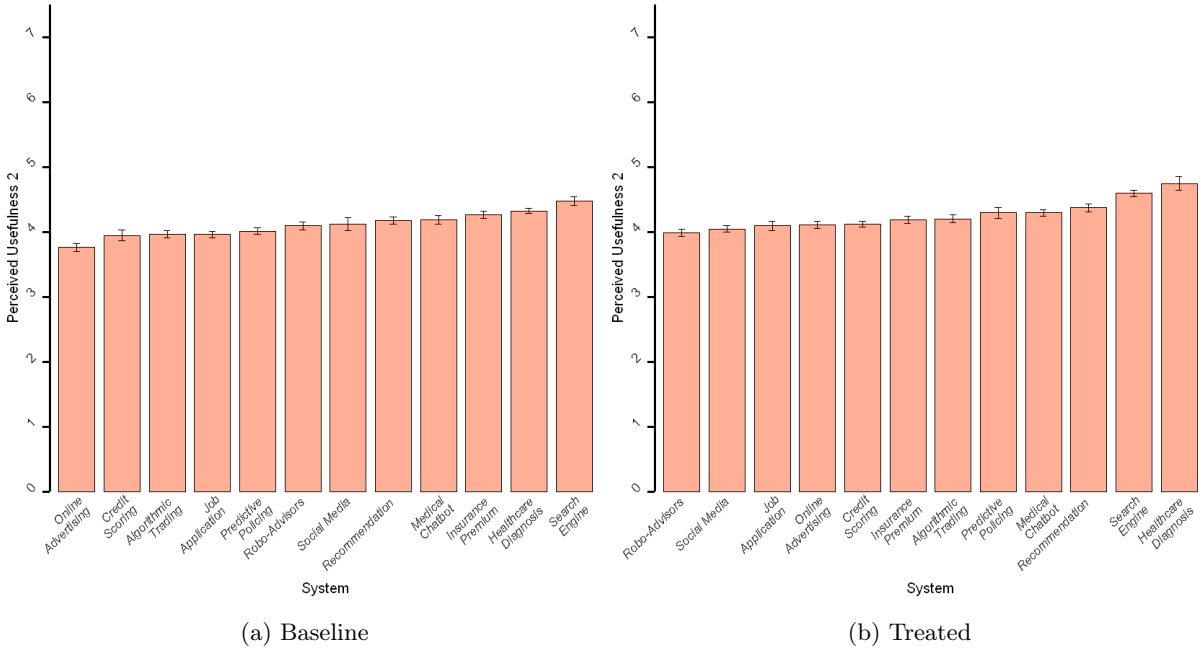


Figure 4.A.17: Willingness to Delegate (Weak Delegation, WTD1)

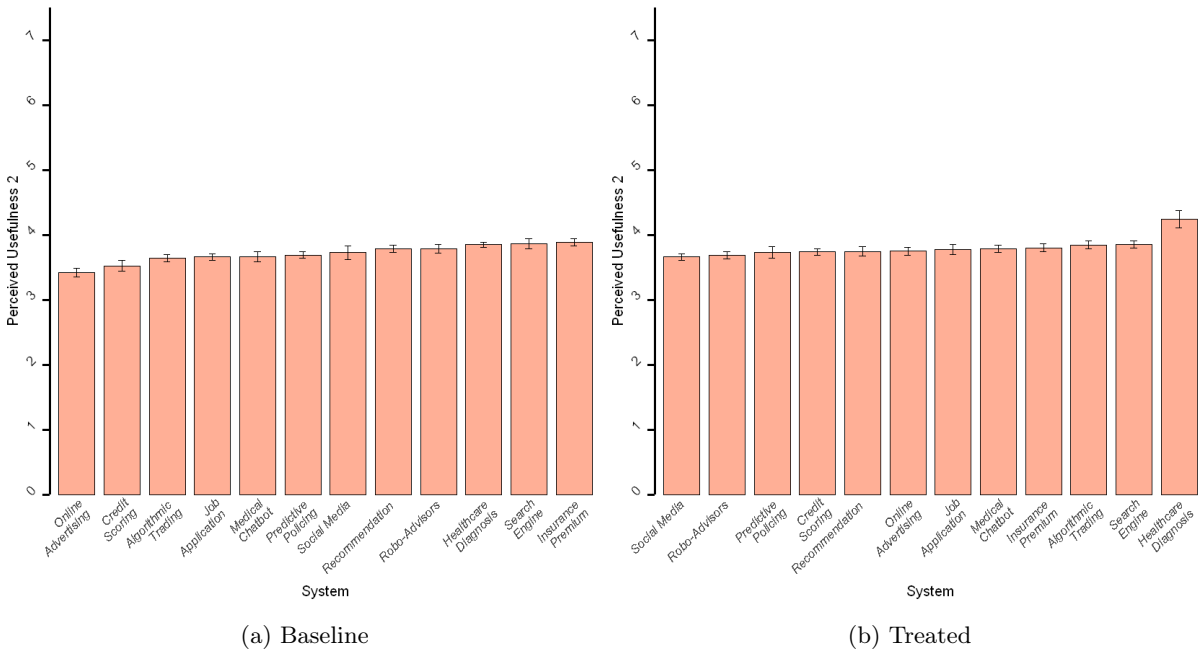


Figure 4.A.18: Willingness to Delegate (Strong Delegation, WTD2)

4.8.2 Descriptive Statistics

Sample

Table 4.A.4: Sample characteristics by country

	France	Germany	Italy
N	1000	1003	1004
Age			
Mean (SD)	48.09 (17.03)	48.29 (16.67)	47.39 (15.10)
Range [min, max]	[18, 83]	[18, 88]	[18, 86]
Education			
No formal education	12 (1.2%)	8 (0.8%)	2 (0.2%)
Primary education	21 (2.1%)	30 (3.0%)	37 (3.7%)
Secondary education (e.g., high school)	239 (23.9%)	163 (16.3%)	353 (35.2%)
Vocational or technical training	189 (18.9%)	307 (30.6%)	152 (15.1%)
Associate degree or equivalent	197 (19.7%)	177 (17.6%)	86 (8.6%)
Bachelor's degree	158 (15.8%)	126 (12.6%)	119 (11.9%)
Master's degree	141 (14.1%)	156 (15.6%)	192 (19.1%)
Doctorate or equivalent	35 (3.5%)	29 (2.9%)	60 (6.0%)
Prefer not to say	8 (0.8%)	7 (0.7%)	3 (0.3%)
Employment			
Employed full-time	499 (49.9%)	431 (43.0%)	461 (45.9%)
Employed part-time	72 (7.2%)	139 (13.9%)	80 (8.0%)
Self-employed	34 (3.4%)	58 (5.8%)	106 (10.6%)
Unemployed (looking for work)	40 (4.0%)	31 (3.1%)	68 (6.8%)
Unemployed (not looking for work)	9 (0.9%)	4 (0.4%)	5 (0.5%)
Retired	252 (25.2%)	234 (23.3%)	121 (12.1%)
Student	42 (4.2%)	29 (2.9%)	66 (6.6%)
Homemaker	34 (3.4%)	52 (5.2%)	87 (8.7%)
Unable to work	15 (1.5%)	22 (2.2%)	5 (0.5%)
Prefer not to say	3 (0.3%)	3 (0.3%)	5 (0.5%)

Table 4.A.4: Sample characteristics by country (*continued*)

	France	Germany	Italy
Gender			
Male	479 (47.9%)	482 (48.1%)	479 (47.7%)
Female	519 (51.9%)	514 (51.2%)	522 (52.0%)
Non-binary / Third gender	2 (0.2%)	3 (0.3%)	NA
Prefer to self-describe	NA	3 (0.3%)	NA
Prefer not to say	NA	1 (0.1%)	3 (0.3%)
Income			
No income	17 (1.7%)	27 (2.7%)	77 (7.7%)
Under \$10,000	103 (10.3%)	84 (8.4%)	91 (9.1%)
\$10,000–\$24,999	301 (30.1%)	225 (22.4%)	259 (25.8%)
\$25,000–\$39,999	281 (28.1%)	227 (22.6%)	281 (28.0%)
\$40,000–\$59,999	165 (16.5%)	177 (17.6%)	142 (14.1%)
\$60,000–\$79,999	56 (5.6%)	113 (11.3%)	52 (5.2%)
\$80,000–\$99,999	21 (2.1%)	53 (5.3%)	23 (2.3%)
\$100,000 or more	16 (1.6%)	48 (4.8%)	20 (2.0%)
Prefer not to say	40 (4.0%)	49 (4.9%)	59 (5.9%)
Marital Status			
Single	274 (27.4%)	103 (10.3%)	274 (27.3%)
Never married	8 (0.8%)	192 (19.1%)	21 (2.1%)
Married or in a registered partnership	475 (47.5%)	464 (46.3%)	447 (44.5%)
Living with partner (cohabiting)	125 (12.5%)	116 (11.6%)	156 (15.5%)
Separated; Divorced	93 (9.3%)	92 (9.2%)	77 (7.7%)
Widowed	24 (2.4%)	34 (3.4%)	22 (2.2%)
Prefer not to say	1 (0.1%)	2 (0.2%)	7 (0.7%)

Beliefs Items

Table 4.A.5: Mean belief score per system and overall belief score, Baseline

System	Self-Reported Knowledge M / SD	Self-Reported Familiarity M / SD	Perceived Exposure M / SD	Self-Reported Use M / SD	Cluster
Algorithmic Trading	3.72 (1.87)	3.32 (1.83)	2.61 (1.85)	2.43 (1.83)	3
Credit Scoring	4.01 (1.82)	3.60 (1.77)	2.83 (1.84)	2.52 (1.79)	2
Healthcare Diagnosis	4.07 (1.76)	3.53 (1.75)	2.77 (1.82)	2.52 (1.82)	2
Insurance Premium	4.04 (1.69)	3.79 (1.71)	3.10 (1.77)	2.78 (1.82)	2
Job Application	3.90 (1.81)	3.55 (1.81)	2.74 (1.90)	2.55 (1.90)	2
Medical Chatbot	4.11 (1.74)	3.53 (1.80)	2.64 (1.81)	2.56 (1.82)	2
Online Advertising	4.31 (1.80)	4.03 (1.83)	4.18 (2.14)	3.43 (2.16)	4
Predictive Policing	3.43 (1.85)	3.07 (1.77)	2.36 (1.83)	2.16 (1.76)	3
Recommendation	4.35 (1.75)	4.12 (1.80)	3.95 (2.12)	3.55 (2.12)	4
Robo-Advisors	3.85 (1.80)	3.56 (1.77)	2.87 (1.91)	2.58 (1.89)	2
Search Engine	4.52 (1.65)	4.46 (1.64)	4.48 (1.98)	4.31 (2.03)	1
Social Media	4.55 (1.66)	4.33 (1.67)	4.40 (1.99)	4.13 (2.06)	1
Mean	4.07 (1.77)	3.74 (1.76)	3.24 (1.91)	2.96 (1.92)	/

Table 4.A.6: Mean belief score per system and overall belief score, Treated

System	Self-Reported Knowledge M / SD	Self-Reported Familiarity M / SD	Perceived Exposure M / SD	Self-Reported Use M / SD	Cluster
Algorithmic Trading	4.00 (1.80)	3.55 (1.73)	2.92 (2.08)	2.61 (1.99)	3
Credit Scoring	3.95 (1.70)	3.69 (1.69)	2.86 (1.86)	2.64 (1.84)	3
Healthcare Diagnosis	4.36 (1.57)	4.17 (1.59)	3.03 (1.82)	3.03 (1.97)	4
Insurance Premium	3.83 (1.79)	3.62 (1.75)	2.80 (1.78)	2.56 (1.83)	3
Job Application	3.93 (1.82)	3.43 (1.80)	2.53 (1.89)	2.37 (1.90)	3
Medical Chatbot	4.32 (1.59)	3.81 (1.66)	2.97 (1.87)	2.83 (1.87)	4
Online Advertising	4.66 (1.54)	4.44 (1.53)	4.95 (1.88)	4.38 (2.04)	1
Predictive Policing	3.64 (1.84)	3.17 (1.81)	2.37 (1.81)	2.16 (1.77)	3
Recommendation	4.74 (1.53)	4.44 (1.56)	4.43 (1.90)	4.13 (1.97)	1
Robo-Advisors	3.80 (1.88)	3.33 (1.83)	2.48 (1.84)	2.39 (1.86)	3
Search Engine	4.79 (1.60)	4.59 (1.61)	4.75 (1.98)	4.63 (2.02)	1
Social Media	4.59 (1.68)	4.35 (1.69)	4.46 (2.17)	4.10 (2.22)	2
Mean	4.22 (1.69)	3.88 (1.69)	3.38 (1.91)	3.15 (1.94)	/

Attitudes Items

Table 4.A.7: Mean attitude score per system and overall attitude score, Baseline (Panel A: Transparency, Fairness, Accountability)

System	TC1	TC2	PF1	PF2	PA1	PA2	Cluster
	M / SD	M / SD	M / SD	M / SD	M / SD	M / SD	
Algorithmic	3.47	3.37	4.09	4.09	4.00	3.84	1
Trading	(1.48)	(1.49)	(1.49)	(1.49)	(1.47)	(1.57)	
Credit	3.37	3.19	4.08	4.09	4.00	3.71	3
Scoring	(1.50)	(1.56)	(1.56)	(1.52)	(1.57)	(1.68)	
Healthcare	3.42	3.20	4.37	4.41	4.29	4.03	4
Diagnosis	(1.43)	(1.42)	(1.46)	(1.44)	(1.47)	(1.55)	
Insurance	3.37	3.24	4.38	4.33	4.28	4.03	4
Premium	(1.38)	(1.43)	(1.41)	(1.41)	(1.41)	(1.55)	
Job	3.42	3.27	4.05	4.09	4.04	3.88	1
Application	(1.49)	(1.49)	(1.52)	(1.51)	(1.51)	(1.61)	
Medical	3.50	3.34	4.31	4.25	4.12	3.83	1
Chatbot	(1.41)	(1.50)	(1.43)	(1.49)	(1.51)	(1.57)	
Online	3.44	3.37	3.93	3.86	3.86	3.66	3
Advertising	(1.51)	(1.53)	(1.51)	(1.58)	(1.59)	(1.66)	
Predictive	3.44	3.29	4.15	4.06	4.08	3.90	1
Policing	(1.51)	(1.51)	(1.51)	(1.52)	(1.50)	(1.54)	
Recommendation	3.42	3.32	4.30	4.27	4.13	3.93	4
	(1.43)	(1.42)	(1.42)	(1.48)	(1.52)	(1.58)	
Robo-Advisors	3.35	3.22	4.25	4.23	4.12	3.95	4
	(1.34)	(1.46)	(1.42)	(1.40)	(1.45)	(1.52)	
Search Engine	3.46	3.25	4.45	4.37	4.15	3.92	2
	(1.36)	(1.33)	(1.36)	(1.37)	(1.38)	(1.48)	
Social Media	3.47	3.41	4.11	4.14	3.96	3.94	1
	(1.44)	(1.49)	(1.48)	(1.52)	(1.52)	(1.73)	
Mean	3.43	3.29	4.21	4.18	4.08	3.89	
	(1.44)	(1.47)	(1.46)	(1.48)	(1.49)	(1.59)	

Table 4.A.8: Mean attitude score per system and overall attitude score, Baseline (Panel B: Privacy, Trust, Use)

System	PC1	PC2	PTru1	PTru2	PU1	PU2	Cluster
	M / SD	M / SD	M / SD	M / SD	M / SD	M / SD	
Algorithmic	3.42	4.00	3.98	4.21	3.87	3.97	1
Trading	(1.51)	(1.55)	(1.53)	(1.47)	(1.53)	(1.54)	
Credit	3.35	3.99	4.00	4.19	3.81	3.81	3
Scoring	(1.53)	(1.57)	(1.62)	(1.43)	(1.59)	(1.64)	
Healthcare	3.46	4.27	4.39	4.58	4.23	4.25	4
Diagnosis	(1.55)	(1.53)	(1.46)	(1.41)	(1.49)	(1.51)	
Insurance	3.43	4.25	4.23	4.41	4.15	4.19	4
Premium	(1.43)	(1.48)	(1.43)	(1.39)	(1.44)	(1.48)	
Job	3.48	4.01	3.96	4.27	3.84	3.93	1
Application	(1.58)	(1.56)	(1.55)	(1.48)	(1.60)	(1.60)	
Medical	3.44	4.04	4.15	4.38	4.00	4.05	1
Chatbot	(1.56)	(1.55)	(1.45)	(1.42)	(1.50)	(1.46)	
Online	3.16	3.68	3.78	4.16	3.75	3.77	3
Advertising	(1.57)	(1.62)	(1.63)	(1.54)	(1.62)	(1.67)	
Predictive	3.36	3.99	4.10	4.29	3.96	3.96	1
Policing	(1.54)	(1.57)	(1.51)	(1.49)	(1.52)	(1.55)	
Recommendation	3.29	4.05	4.24	4.45	4.10	4.19	4
	(1.49)	(1.57)	(1.50)	(1.44)	(1.49)	(1.52)	
Robo-Advisors	3.45	4.17	4.19	4.37	4.09	4.19	4
	(1.47)	(1.47)	(1.44)	(1.43)	(1.47)	(1.48)	
Search Engine	3.17	4.17	4.39	4.49	4.30	4.27	2
	(1.43)	(1.41)	(1.36)	(1.34)	(1.37)	(1.48)	
Social Media	3.22	3.99	4.05	4.14	4.03	4.04	1
	(1.48)	(1.63)	(1.48)	(1.39)	(1.55)	(1.55)	
Mean	3.35	4.05	4.12	4.33	4.01	4.05	
	(1.51)	(1.54)	(1.50)	(1.44)	(1.51)	(1.54)	

Table 4.A.9: Mean attitude score per system and overall attitude score, Treated (Panel A: Transparency, Fairness, Accountability)

System	TC1	TC2	PF1	PF2	PA1	PA2	Cluster
	M / SD	M / SD	M / SD	M / SD	M / SD	M / SD	
Algorithmic	3.38	3.28	4.31	4.40	4.18	4.02	4
Trading	(1.39)	(1.42)	(1.45)	(1.43)	(1.50)	(1.61)	
Credit	3.42	3.25	4.26	4.26	4.13	3.95	4
Scoring	(1.40)	(1.39)	(1.44)	(1.44)	(1.49)	(1.57)	
Healthcare	3.56	3.22	4.68	4.78	4.52	4.42	1
Diagnosis	(1.41)	(1.30)	(1.24)	(1.26)	(1.46)	(1.48)	
Insurance	3.39	3.13	4.34	4.30	4.16	4.01	4
Premium	(1.43)	(1.44)	(1.48)	(1.47)	(1.53)	(1.59)	
Job	3.45	3.29	4.18	4.15	3.98	3.82	2
Application	(1.38)	(1.42)	(1.50)	(1.50)	(1.57)	(1.71)	
Medical	3.34	3.16	4.47	4.39	4.32	4.10	4
Chatbot	(1.39)	(1.38)	(1.43)	(1.45)	(1.48)	(1.57)	
Online	3.32	3.25	4.24	4.22	4.16	3.93	2
Advertising	(1.34)	(1.30)	(1.47)	(1.48)	(1.52)	(1.59)	
Predictive	3.67	3.35	4.27	4.30	4.19	3.93	3
Policing	(1.50)	(1.48)	(1.55)	(1.52)	(1.65)	(1.67)	
Recommendation	3.50	3.23	4.42	4.37	4.16	3.91	4
	(1.40)	(1.34)	(1.41)	(1.48)	(1.51)	(1.65)	
Robo-Advisors	3.49	3.32	4.23	4.24	4.07	3.80	3
	(1.51)	(1.48)	(1.52)	(1.52)	(1.57)	(1.70)	
Search Engine	3.49	3.26	4.53	4.50	4.28	4.01	4
	(1.41)	(1.36)	(1.40)	(1.41)	(1.50)	(1.60)	
Social Media	3.38	3.20	4.20	4.16	4.09	3.83	2
	(1.43)	(1.39)	(1.47)	(1.51)	(1.55)	(1.66)	
Mean	3.45	3.25	4.34	4.34	4.19	3.98	
	(1.42)	(1.39)	(1.45)	(1.46)	(1.53)	(1.62)	

Table 4.A.10: Mean attitude score per system and overall attitude score, Treated (Panel B: Privacy, Trust, Use)

System	PC1	PC2	PTru1	PTru2	PU1	PU2	Cluster
	M / SD	M / SD	M / SD	M / SD	M / SD	M / SD	
Algorithmic	3.30	4.25	4.31	4.43	4.14	4.26	4
Trading	(1.47)	(1.55)	(1.46)	(1.40)	(1.54)	(1.52)	
Credit	3.36	4.18	4.26	4.35	4.04	3.99	4
Scoring	(1.49)	(1.52)	(1.47)	(1.41)	(1.47)	(1.52)	
Healthcare	3.56	4.68	4.72	4.89	4.70	4.79	1
Diagnosis	(1.49)	(1.32)	(1.26)	(1.25)	(1.30)	(1.26)	
Insurance	3.33	4.11	4.25	4.43	4.07	4.08	4
Premium	(1.53)	(1.52)	(1.51)	(1.43)	(1.53)	(1.54)	
Job	3.49	4.12	4.10	4.31	3.93	3.95	2
Application	(1.53)	(1.54)	(1.51)	(1.41)	(1.54)	(1.58)	
Medical	3.36	4.24	4.31	4.60	4.09	4.14	4
Chatbot	(1.52)	(1.54)	(1.49)	(1.37)	(1.52)	(1.57)	
Online	3.17	4.08	4.26	4.43	4.04	4.09	2
Advertising	(1.43)	(1.61)	(1.53)	(1.40)	(1.51)	(1.53)	
Predictive	3.58	4.13	4.22	4.48	4.18	4.05	3
Policing	(1.59)	(1.66)	(1.61)	(1.52)	(1.57)	(1.63)	
Recommendation	3.37	4.21	4.32	4.55	4.24	4.30	4
	(1.53)	(1.54)	(1.51)	(1.44)	(1.47)	(1.55)	
Robo-Advisors	3.30	4.06	4.14	4.33	4.06	4.07	3
	(1.62)	(1.62)	(1.57)	(1.51)	(1.55)	(1.62)	
Search Engine	3.33	4.31	4.51	4.65	4.37	4.45	4
	(1.48)	(1.52)	(1.44)	(1.41)	(1.46)	(1.51)	
Social Media	3.18	3.98	4.17	4.39	4.03	4.06	2
	(1.50)	(1.62)	(1.55)	(1.43)	(1.57)	(1.62)	
Mean	3.36	4.19	4.30	4.49	4.16	4.19	
	(1.51)	(1.54)	(1.49)	(1.42)	(1.50)	(1.54)	

Willingness to Delegate Items

Table 4.A.11: Mean WTD score per system and overall WTD score, Baseline

System	Willingness to Delegate (Weak)	Willingness to Delegate (Strong)	Cluster
	M / SD	M / SD	
Algorithmic Trading	3.96 (1.58)	3.64 (1.65)	2
Credit Scoring	3.95 (1.65)	3.52 (1.73)	4
Healthcare Diagnosis	4.32 (1.53)	3.85 (1.61)	1
Insurance Premium	4.27 (1.49)	3.89 (1.57)	1
Job Application	3.96 (1.62)	3.66 (1.66)	2
Medical Chatbot	4.18 (1.50)	3.66 (1.67)	3
Online Advertising	3.76 (1.72)	3.42 (1.74)	4
Predictive Policing	4.01 (1.58)	3.69 (1.67)	2
Recommendation	4.17 (1.56)	3.78 (1.66)	3
Robo-Advisors	4.09 (1.53)	3.79 (1.61)	3
Search Engine	4.47 (1.42)	3.86 (1.63)	1
Social Media	4.12 (1.60)	3.72 (1.70)	2
Mean	4.11 (1.57)	3.71 (1.66)	

Table 4.A.12: Mean WTD score per system and overall WTD score, Treated

System	Willingness to Delegate (Weak)	Willingness to Delegate (Strong)	Cluster
Algorithmic Trading	4.20 (1.55)	3.85 (1.64)	4
Credit Scoring	4.12 (1.54)	3.74 (1.61)	4
Healthcare Diagnosis	4.74 (1.33)	4.24 (1.60)	1
Insurance Premium	4.19 (1.55)	3.80 (1.70)	2
Job Application	4.09 (1.59)	3.77 (1.68)	4
Medical Chatbot	4.30 (1.56)	3.79 (1.73)	2
Online Advertising	4.11 (1.60)	3.75 (1.73)	3
Predictive Policing	4.29 (1.59)	3.73 (1.71)	2
Recommendation	4.37 (1.53)	3.75 (1.72)	2
Robo-Advisors	3.99 (1.70)	3.68 (1.74)	3
Search Engine	4.59 (1.49)	3.85 (1.68)	2
Social Media	4.05 (1.63)	3.66 (1.71)	3
Mean	4.25 (1.55)	3.80 (1.69)	

4.8.3 Additional Tables

Table 4.A.13: WTD1 decomposition by system: Sum indices vs PCA indices

Component	Sum indices	PCA indices
<i>Algorithmic Trading</i>		
Direct ($\hat{d}$)	0.031 (0.033)	0.029 (0.034)
Indirect via beliefs ($\hat{a}\hat{e}$)	0.023*** (0.008)	0.015*** (0.006)
Indirect via attitudes ($\hat{c}\hat{f}$)	0.061* (0.032)	0.059* (0.032)
Indirect via beliefs→attitudes ($\hat{a}\hat{b}\hat{f}$)	0.054*** (0.018)	0.066*** (0.022)
Total	0.169*** (0.054)	0.169*** (0.054)
<i>Credit Scoring</i>		
Direct ($\hat{d}$)	-0.010 (0.040)	0.021 (0.043)
Indirect via beliefs ($\hat{a}\hat{e}$)	0.012 (0.009)	0.007 (0.006)
Indirect via attitudes ($\hat{c}\hat{f}$)	0.064 (0.038)	0.036 (0.036)
Indirect via beliefs→attitudes ($\hat{a}\hat{b}\hat{f}$)	0.024 (0.018)	0.028 (0.021)
Total	0.091 (0.057)	0.091 (0.057)
<i>Healthcare Diagnosis</i>		
Direct ($\hat{d}$)	-0.012 (0.055)	0.039 (0.057)
Indirect via beliefs ($\hat{a}\hat{e}$)	0.018 (0.011)	0.007* (0.005)
Indirect via attitudes ($\hat{c}\hat{f}$)	0.156*** (0.043)	0.094** (0.042)
Indirect via beliefs→attitudes ($\hat{a}\hat{b}\hat{f}$)	0.040 (0.024)	0.063* (0.032)
Total	0.203** (0.074)	0.203** (0.074)
<i>Insurance Premium</i>		
Direct ($\hat{d}$)	0.047 (0.047)	0.000 (0.046)
Indirect via beliefs ($\hat{a}\hat{e}$)	0.015 (0.011)	0.007 (0.006)
Indirect via attitudes ($\hat{c}\hat{f}$)	0.024 (0.039)	0.075* (0.040)
Indirect via beliefs→attitudes ($\hat{a}\hat{b}\hat{f}$)	0.028 (0.020)	0.033 (0.025)
Total	0.115* (0.067)	0.115* (0.067)
<i>Job Application</i>		
Direct ($\hat{d}$)	0.067* (0.038)	0.087** (0.039)
Indirect via beliefs ($\hat{a}\hat{e}$)	0.010 (0.009)	0.006 (0.005)
Indirect via attitudes ($\hat{c}\hat{f}$)	0.083** (0.034)	0.056* (0.033)
Indirect via beliefs→attitudes ($\hat{a}\hat{b}\hat{f}$)	0.021 (0.018)	0.031 (0.023)
Total	0.181*** (0.057)	0.181*** (0.057)
<i>Medical Chatbot</i>		
Direct ($\hat{d}$)	0.001 (0.036)	-0.026 (0.037)
Indirect via beliefs ($\hat{a}\hat{e}$)	0.011* (0.007)	0.007* (0.005)

WTD1 decomposition by system (continued)

Component	Sum indices	PCA indices
Indirect via attitudes ($\hat{c}\hat{f}$)	0.025 (0.034)	0.051 (0.035)
Indirect via beliefs→attitudes ($\hat{a}\hat{b}\hat{f}$)	0.031* (0.019)	0.037* (0.022)
Total	0.068 (0.056)	0.068 (0.056)
<i>Online Advertising</i>		
Direct ($\hat{d}$)	-0.009 (0.040)	-0.025 (0.041)
Indirect via beliefs ($\hat{a}\hat{e}$)	0.029*** (0.010)	0.004 (0.004)
Indirect via attitudes ($\hat{c}\hat{f}$)	0.097** (0.047)	0.117*** (0.046)
Indirect via beliefs→attitudes ($\hat{a}\hat{b}\hat{f}$)	0.041*** (0.014)	0.062** (0.022)
Total	0.158** (0.067)	0.158** (0.067)
<i>Predictive Policing</i>		
Direct ($\hat{d}$)	0.078* (0.042)	0.124*** (0.044)
Indirect via beliefs ($\hat{a}\hat{e}$)	0.007 (0.007)	0.005 (0.004)
Indirect via attitudes ($\hat{c}\hat{f}$)	0.165*** (0.049)	0.111** (0.044)
Indirect via beliefs→attitudes ($\hat{a}\hat{b}\hat{f}$)	0.021 (0.021)	0.030 (0.024)
Total	0.270*** (0.069)	0.270*** (0.069)
<i>Recommendation</i>		
Direct ($\hat{d}$)	0.001 (0.035)	0.020 (0.036)
Indirect via beliefs ($\hat{a}\hat{e}$)	0.050*** (0.012)	0.030*** (0.009)
Indirect via attitudes ($\hat{c}\hat{f}$)	0.003 (0.035)	-0.015 (0.035)
Indirect via beliefs→attitudes ($\hat{a}\hat{b}\hat{f}$)	0.059*** (0.014)	0.078*** (0.018)
Total	0.113** (0.053)	0.113** (0.053)
<i>Robo-Advisors</i>		
Direct ($\hat{d}$)	0.032 (0.045)	0.025 (0.044)
Indirect via beliefs ($\hat{a}\hat{e}$)	-0.003 (0.009)	-0.002 (0.008)
Indirect via attitudes ($\hat{c}\hat{f}$)	0.007 (0.052)	0.012 (0.044)
Indirect via beliefs→attitudes ($\hat{a}\hat{b}\hat{f}$)	-0.008 (0.027)	-0.007 (0.028)
Total	0.028 (0.070)	0.028 (0.070)
<i>Search Engine</i>		
Direct ($\hat{d}$)	-0.010 (0.039)	0.022 (0.040)
Indirect via beliefs ($\hat{a}\hat{e}$)	-0.005 (0.013)	-0.003 (0.009)
Indirect via attitudes ($\hat{c}\hat{f}$)	0.056* (0.032)	0.026 (0.032)
Indirect via beliefs→attitudes ($\hat{a}\hat{b}\hat{f}$)	-0.004 (0.011)	-0.007 (0.017)
Total	0.038 (0.055)	0.038 (0.055)
<i>Social Media</i>		

WTD1 decomposition by system (continued)

Component	Sum indices	PCA indices
Direct ($\hat{d}$)	-0.049 (0.047)	-0.070 (0.049)
Indirect via beliefs ($\hat{a}\hat{e}$)	0.005 (0.013)	0.004 (0.006)
Indirect via attitudes ($\hat{c}\hat{f}$)	0.065 (0.048)	0.080* (0.049)
Indirect via beliefs→attitudes ($\hat{a}\hat{b}\hat{f}$)	0.004 (0.011)	0.011 (0.018)
Total	0.025 (0.070)	0.025 (0.070)

Table 4.A.14: WTD2 decomposition by system: Sum indices vs PCA indices

Component	Sum indices	PCA indices
<i>Algorithmic Trading</i>		
Direct ($\hat{d}$)	-0.042 (0.035)	-0.036 (0.038)
Indirect via beliefs ($\hat{a}\hat{e}$)	0.024*** (0.009)	0.021*** (0.008)
Indirect via attitudes ($\hat{c}\hat{f}$)	0.057* (0.030)	0.049* (0.027)
Indirect via beliefs→attitudes ($\hat{a}\hat{b}\hat{f}$)	0.050*** (0.017)	0.055*** (0.018)
Total	0.089* (0.052)	0.089* (0.052)
<i>Credit Scoring</i>		
Direct ($\hat{d}$)	-0.009 (0.038)	0.024 (0.041)
Indirect via beliefs ($\hat{a}\hat{e}$)	0.011 (0.008)	0.007 (0.006)
Indirect via attitudes ($\hat{c}\hat{f}$)	0.061 (0.036)	0.031 (0.032)
Indirect via beliefs→attitudes ($\hat{a}\hat{b}\hat{f}$)	0.023 (0.017)	0.024 (0.019)
Total	0.086 (0.054)	0.086 (0.054)
<i>Healthcare Diagnosis</i>		
Direct ($\hat{d}$)	-0.037 (0.065)	0.017 (0.062)
Indirect via beliefs ($\hat{a}\hat{e}$)	0.023 (0.014)	0.017* (0.009)
Indirect via attitudes ($\hat{c}\hat{f}$)	0.138*** (0.038)	0.076** (0.034)
Indirect via beliefs→attitudes ($\hat{a}\hat{b}\hat{f}$)	0.035 (0.021)	0.050* (0.025)
Total	0.159* (0.082)	0.159* (0.082)
<i>Insurance Premium</i>		
Direct ($\hat{d}$)	-0.009 (0.049)	-0.050 (0.052)
Indirect via beliefs ($\hat{a}\hat{e}$)	0.013 (0.009)	0.007 (0.006)
Indirect via attitudes ($\hat{c}\hat{f}$)	0.024 (0.038)	0.068* (0.036)
Indirect via beliefs→attitudes ($\hat{a}\hat{b}\hat{f}$)	0.027 (0.019)	0.030 (0.023)
Total	0.054 (0.069)	0.054 (0.069)
<i>Job Application</i>		
Direct ($\hat{d}$)	0.024 (0.040)	0.047 (0.041)

WTD2 decomposition by system (continued)

Component	Sum indices	PCA indices
Indirect via beliefs ($\hat{a}\hat{e}$)	0.010 (0.009)	0.009 (0.007)
Indirect via attitudes ($\hat{c}\hat{f}$)	0.076** (0.031)	0.047* (0.028)
Indirect via beliefs→attitudes ($\hat{a}\hat{b}\hat{f}$)	0.019 (0.017)	0.026 (0.019)
Total	0.130** (0.057)	0.130** (0.057)
<i>Medical Chatbot</i>		
Direct ($\hat{d}$)	0.006 (0.041)	-0.012 (0.042)
Indirect via beliefs ($\hat{a}\hat{e}$)	0.017* (0.010)	0.013* (0.008)
Indirect via attitudes ($\hat{c}\hat{f}$)	0.022 (0.030)	0.041 (0.028)
Indirect via beliefs→attitudes ($\hat{a}\hat{b}\hat{f}$)	0.027* (0.016)	0.030* (0.018)
Total	0.072 (0.056)	0.072 (0.056)
<i>Online Advertising</i>		
Direct ($\hat{d}$)	0.042 (0.043)	0.025 (0.043)
Indirect via beliefs ($\hat{a}\hat{e}$)	0.002 (0.004)	-0.019** (0.008)
Indirect via attitudes ($\hat{c}\hat{f}$)	0.093** (0.045)	0.111*** (0.044)
Indirect via beliefs→attitudes ($\hat{a}\hat{b}\hat{f}$)	0.039*** (0.013)	0.059** (0.021)
Total	0.177*** (0.059)	0.177*** (0.059)
<i>Predictive Policing</i>		
Direct ($\hat{d}$)	-0.061 (0.044)	-0.012 (0.050)
Indirect via beliefs ($\hat{a}\hat{e}$)	0.010 (0.011)	0.012 (0.010)
Indirect via attitudes ($\hat{c}\hat{f}$)	0.144*** (0.043)	0.089** (0.036)
Indirect via beliefs→attitudes ($\hat{a}\hat{b}\hat{f}$)	0.018 (0.018)	0.024 (0.019)
Total	0.112 (0.067)	0.112 (0.067)
<i>Recommendation</i>		
Direct ($\hat{d}$)	-0.046 (0.042)	-0.028 (0.043)
Indirect via beliefs ($\hat{a}\hat{e}$)	-0.005 (0.005)	-0.025*** (0.008)
Indirect via attitudes ($\hat{c}\hat{f}$)	0.003 (0.037)	-0.016 (0.037)
Indirect via beliefs→attitudes ($\hat{a}\hat{b}\hat{f}$)	0.062*** (0.014)	0.083*** (0.019)
Total	0.014 (0.058)	0.014 (0.058)
<i>Robo-Advisors</i>		
Direct ($\hat{d}$)	-0.016 (0.049)	-0.022 (0.055)
Indirect via beliefs ($\hat{a}\hat{e}$)	-0.004 (0.013)	-0.003 (0.012)
Indirect via attitudes ($\hat{c}\hat{f}$)	0.006 (0.043)	0.009 (0.034)
Indirect via beliefs→attitudes ($\hat{a}\hat{b}\hat{f}$)	-0.007 (0.022)	-0.005 (0.022)
Total	-0.021 (0.074)	-0.021 (0.074)

WTD2 decomposition by system (continued)

Component	Sum indices	PCA indices
<i>Search Engine</i>		
Direct ($\hat{d}$)	-0.036 (0.045)	-0.000 (0.048)
Indirect via beliefs ($\hat{a}\hat{e}$)	0.002 (0.004)	0.004 (0.009)
Indirect via attitudes ($\hat{c}\hat{f}$)	0.065* (0.037)	0.030 (0.037)
Indirect via beliefs→attitudes ($\hat{a}\hat{b}\hat{f}$)	-0.005 (0.013)	-0.008 (0.019)
Total	0.026 (0.061)	0.026 (0.061)
<i>Social Media</i>		
Direct ($\hat{d}$)	-0.004 (0.047)	-0.022 (0.049)
Indirect via beliefs ($\hat{a}\hat{e}$)	-0.000 (0.001)	-0.004 (0.008)
Indirect via attitudes ($\hat{c}\hat{f}$)	0.068 (0.050)	0.084* (0.051)
Indirect via beliefs→attitudes ($\hat{a}\hat{b}\hat{f}$)	0.004 (0.011)	0.011 (0.019)
Total	0.068 (0.072)	0.068 (0.072)

General Conclusion

General Conclusion

Over the last decade, the production and circulation of individual-level data have become a core input to economic activity. Households and consumers generate dense “behavioral exhaust” (financial transactions, browsing histories, and social interactions) which are collected, processed, and transformed into predictions and recommendations. In parallel, algorithmic systems have become decision intermediaries: they classify, score, rank, match, and nudge. This shift has created a new set of economic frictions that standard models of information and incentives only partially capture. Privacy is no longer a peripheral constraint but a strategic and behavioral margin of choice, and algorithmic opacity has become a determinant of trust, adoption, and the legitimacy of delegation.

This dissertation examined these frictions through four essays on privacy risk and algorithmic transparency. The unifying objective was to provide empirically grounded evidence on how individuals *value* privacy risk, how they *respond* to algorithmic explanations, and under what conditions transparency policies can plausibly improve welfare. Across contexts—financial data leaks, robo-advising, music recommendation, and public acceptability of algorithmic decision-making (ADM)—the dissertation treats digital mediation as an *information environment*: a structured set of incentives, beliefs, attention constraints, and social signals that jointly determine behavior.

A key conclusion across the four essays is that improving the information environment is not equivalent to changing behavior. Transparency and disclosure alter beliefs and perceived understanding, but their behavioral impact is bounded by the underlying payoff structure and by endogenous information acquisition and processing costs. When privacy risks are made explicit and immediate, individuals do respond in the expected direction, but incentives are typically more salient than probabilistic harms. When algorithmic explanations are provided, they often improve perceived understanding and reshape beliefs, yet average behavioral effects are limited and strongly heterogeneous. Transparency, therefore, matters, but it is not sufficient on its own. Whether it changes choices depends on attention, context of use, and the extent to which users actively process the information. Interpreted through the lens of information economics, transparency and disclosure act as interventions on the *information structure* facing agents. However, their welfare and behavioral effects are bounded when attention and processing are costly and

optimally allocated. Put differently, additional information can improve perceived comprehension and posterior beliefs, while producing only second-order changes in choices unless it alters the effective payoff or the constraint set. This framing aligns with economic perspectives that emphasize limited attention and rational inattention in complex choice environments (Sims, 2003; Matějka & McKay, 2015).

The remainder of this conclusion synthesizes the main findings of each essay, discusses the main contributions of this thesis, the policy implications, and outlines promising avenues for future research.

Summary of the four essays

The first essay adopts an incentivized experimental methodology to elicit privacy–monetary trade-offs in a setting where privacy risk is unusually transparent and consequential. Participants repeatedly choose between alternatives that differ in expected monetary returns and in the probability of a financial-data leak; probabilities are explicitly displayed, the decision environment is time-varying, and leak outcomes are realized with feedback. The design is intentionally favorable to a standard “privacy calculus” view of behavior. Agents can map choices into expected costs and benefits and it contrasts with many real-world privacy environments where risks are ambiguous, delayed, and embedded in opaque data flows (Dinev & Hart, 2006; Acquisti et al., 2016). Although the approach is primarily empirical, it speaks directly to theory by providing a clean testbed for how individuals process explicit probabilistic harms and how they update from experience and from social information.

Three results summarize the evidence. First, choices respond to both return spreads and leak probabilities in the predicted directions, but responsiveness to returns is substantially larger than responsiveness to privacy risk, even when the privacy dimension is made salient. Second, learning depends on the information channel: own leak experiences have modest and short-lived effects, while exposure to others’ leaks generates a stronger and more persistent shift away from the risky option, consistent with social-information channels and informational cascades (Banerjee, 1992; Bikhchandani et al., 1992). Third, standard preference proxies explain little of the heterogeneity in privacy-risk choices, whereas context-relevant privacy attitudes are more predictive, echoing the privacy paradox literature (Berendt et al., 2005; Norberg et al., 2007; Dienlin & Trepte, 2015). From a public policy perspective, the benchmark interpretation is informative: if privacy receives limited weight even when risk is explicit and immediate, then purely informational tools (e.g., notice-and-consent) may be insufficient in environments where harms are less visible, strengthening the case for complementary instruments that change incentives, defaults, and security obligations.

The second essay moves from a controlled benchmark to a high-stakes applied setting: robo-advising. Methodologically, it evaluates the effect of explanation provision within a digital advisory process by

varying the transparency of the recommendation and measuring subsequent adherence and deviations. The theoretical motivation is rooted in delegation and information problems: households outsource complex portfolio choices to an algorithmic intermediary, and explanations are often viewed as a tool to reduce informational asymmetries, improve perceived legitimacy, and mitigate potentially conflicting recommendations (D’Acunto & Rossi, 2023; Cardillo & Chiappini, 2024). The context is crucial because robo-advisors are increasingly used for long-horizon wealth accumulation decisions, including retirement savings and pension-related allocations, where small differences in risk exposure and fees can compound into meaningful welfare consequences (Dominitz & Hung, 2006; Hung & Yoong, 2013; Philippon, 2019).

Empirically, the results point to a nuanced conclusion. On average, providing explanations does not substantially change acceptance of the recommended risk level nor the propensity to deviate from the recommended portfolio. The effects are instead concentrated in heterogeneity: behavior depends on the digital context (e.g., device constraints) and on prior user experience with the platform (e.g., first-time users), consistent with evidence that information display, cognitive load, and usage conditions shape portfolio decisions and the use of financial information (Agnew & Szykman, 2005; Deo & Sontakke, 2021; Mograbi, 2022). The public policy implication is that transparency requirements in digital advice should be assessed in the environments in which households actually make retirement-relevant decisions. If explanations primarily shift perceptions for some users while leaving average adherence unchanged, then consumer protection may require combining explainability with design standards (clarity, timing, salience), suitability, and oversight mechanisms tailored to long-term household financial security.

The third essay examines algorithmic explanations in a platform market characterized by abundance and high search costs. The starting point is theoretical: recommender systems can be understood as market-design and information devices that reduce search frictions, reshape demand, and reallocate attention (J. Y. Bakos, 1997), with especially high stakes in platform settings where curation can concentrate visibility and influence consumption (Srnicek, 2017; Aguiar & Waldfogel, 2021). At the same time, recommendation processes are opaque and can embed platform objectives, raising concerns about manipulation, market power, and unequal exposure. In this framework, explanations are not merely “extra information”; they potentially change the perceived mapping between past behavior and future recommendations, affecting users’ willingness to explore, their confidence in the system, and the extent to which they attribute outcomes to their own preferences versus platform curation (Cramer et al., 2008; Tintarev & Masthoff, 2015; Y. Zhang & Chen, 2020).

Methodologically, the essay tests these mechanisms by introducing explanations for music recommender systems and measuring discovery behavior, listening intensity, and engagement, building on and speaking to a growing literature on explainable recommendation in general and music recommendation in particular (Schröder & Ghajargar, 2021; Afchar et al., 2022). The core result is again heterogeneous: ex-

planations do not produce large mean shifts in core outcomes, but among participants who meaningfully engage with the explanations, transparency changes behavior in economically relevant ways. Listening time increases, and the probability of reaching a platform-relevant consumption threshold rises. This suggests that explanations can matter through attention and engagement (i.e., whether information is processed) rather than through uniform average effects. The policy implication is twofold: first, transparency can improve user agency when it succeeds in lowering perceived opacity costs; second, because explanations interact with platform prominence and curation, transparency debates should be connected to broader questions about platform power and the distribution of exposure in cultural markets (Srnicek, 2017; Aguiar & Waldfogel, 2021).

The fourth essay provides large-scale causal evidence on how explanations affect public perceptions of algorithmic decision-making (ADM) across domains. Methodologically, it relies on a broad survey-experimental design implemented at scale, allowing systematic comparisons across multiple application areas and enabling precise estimates of heterogeneous responses. The theoretical motivation is that adoption of ADM depends on multiple dimensions (perceived understanding, fairness, accountability, privacy, and trust), and the sign of explanation effects is ambiguous. Explanations may increase perceived comprehension and procedural legitimacy, but they may also reveal trade-offs, uncertainty, or value-laden criteria that reduce confidence (Castelvecchi, 2016; Doshi-Velez & Kim, 2017; Arrieta et al., 2020). More generally, the essay connects explanation effects to classic questions of trust in automation and to documented patterns of algorithm aversion and algorithm appreciation (Jian et al., 2000; Dietvorst et al., 2015; Logg et al., 2019).

The results confirm strong context dependence. Explanations tend to increase perceived understanding, but do not uniformly translate into higher trust or greater willingness to delegate, and the relative salience of privacy versus fairness and accountability concerns varies across contexts. These patterns are consistent with acceptance frameworks in which transparency concerns shape trust and adoption through intermediate perceptions of fairness, accountability, and privacy (Aysolmaz et al., 2023). The public policy implication is that governance tools should not assume that “more explanation” automatically yields legitimacy: explanation mandates need to be calibrated to context-specific concerns and accompanied by accountability and contestability mechanisms (Kroll et al., 2017; Wieringa, 2020). The large-scale nature of the evidence strengthens this conclusion by showing that these patterns are not confined to a narrow user group or a single context, but emerge systematically across a wide set of respondents and settings.

Main Contributions

A first contribution is to show that incentives dominate privacy risk, even when risk is explicit. A striking result of the dissertation is the persistent asymmetry between reactions to incentives and reactions to risk. In the privacy benchmark, leak probabilities matter, but returns matter more. This asymmetry helps reconcile two literatures. On the one hand, privacy economics documents meaningful willingness to pay for privacy and a rich set of privacy-protective behaviors (Acquisti et al., 2016). On the other hand, the privacy paradox literature highlights repeated gaps between concerns and behavior (Berendt et al., 2005; Norberg et al., 2007; Dienlin & Trepte, 2015). The dissertation suggests that one reason for the gap is environmental: when benefits are immediate, quantifiable, and framed as gains, while privacy harms are probabilistic and harder to map to concrete losses, behavior will tend to follow the incentive margin. For policy, the implication is uncomfortable but clear: purely informational interventions that aim to “educate” users about privacy may have limited effects unless they also reshape incentives, defaults, or the timing and salience of consequences.

A second contribution is to argue that transparency is an attention problem before it is a content problem. Across robo-advising and music streaming, the average behavioral effects of explanations are muted. This does not necessarily mean that explanations are ineffective; it may mean that they are not *consumed*. Economic models of limited attention and rational inattention provide a natural interpretation (Sims, 2003; Matějka & McKay, 2015): individuals allocate scarce cognitive resources to information only when its expected decision value exceeds its processing cost. In the settings studied in this dissertation, those processing costs are not abstract. They are shaped by interface design (where explanations are placed and how they are formatted), by time pressure and competing tasks, by device constraints (e.g., mobile versus desktop), and by the user’s familiarity with the service and the domain.

A key methodological contribution of the dissertation is to operationalize this mechanism by measuring engagement and treating transparency as a two-stage process: information must be *provided*, and then it must be *taken up*. This distinction helps reconcile why explanations can systematically move perceived understanding and reported confidence while leaving average choices largely unchanged: the relevant margin is often whether users notice, attend to, and interpret the information at the moment of decision. This perspective also connects naturally to a broader disclosure literature showing that consumers frequently do not read standard-form contracts and disclosures even when they are available (McDonald & Cranor, 2008; Y. Bakos et al., 2014). In short, transparency can fail not because information is inaccurate, but because it is too costly to process relative to its perceived benefits in context.

This contribution also motivates a concrete research agenda. If attention is the binding constraint, then the question for future work is not only *what* to explain, but *how* and *when* to explain. Promising

directions include testing explanation formats that reduce cognitive load (e.g., layered or progressive disclosure), identifying contexts in which explanations have higher marginal value (e.g., high-stakes or unfamiliar decisions), and studying dynamic effects in repeated interactions where users may learn to use explanations more effectively or, conversely, habituate and ignore them. More generally, the results call for evaluating transparency policies as interface-level interventions with heterogeneous treatment effects, rather than as a uniform remedy that automatically translates information into behavior.

A third contribution is to highlight the social foundations of privacy behavior: privacy choices are shaped not only by individual preferences and risk attitudes, but also by what individuals observe about others' experiences and by the narratives that circulate in the public sphere. In the privacy experiment, information about others' leak experiences has stronger and more persistent effects than one's own experience, a pattern consistent with social-information channels and informational cascades (Banerjee, 1992; Bikhchandani et al., 1992). The broader implication is that privacy risk is, in part, a *socially learned* object. Salient breaches, peer experiences, and media coverage can shift perceived prevalence and perceived severity, generating behavioral adjustments that would be hard to predict from stable individual-level risk preferences alone.

This perspective matters for governance because many privacy policies operate through public signals. Breach notification rules, transparency reports, and media coverage can alter behavior, but they may do so by shifting collective salience and shared beliefs rather than by improving private information in a purely individual sense. This creates two practical implications. First, policy evaluations should account for spillovers: the effect of a breach or a disclosure is not confined to those directly harmed, but can propagate through attention and social learning. Second, reliance on salience shocks is fragile: repeated incidents can generate fatigue, uneven updating across groups with different information environments, or overreaction when rare events are made vivid. While the dissertation provides initial causal evidence on these channels, it also points to a second research agenda that deserves to be developed more explicitly.

Future work should map the *networks* through which privacy-relevant information travels (peers, social media, news), identify which signals are most influential (personal stories versus statistics, local versus national events), and quantify how spillovers interact with incentives and platform design. A complementary direction is to study how social learning interacts with transparency: explanations and disclosures may matter more when they are reinforced by social discussion, and less when the prevailing narrative is distrustful or fatalistic. Developing these questions would help connect individual privacy behavior to collective dynamics and would strengthen the policy relevance of privacy regulation in an environment where risk perceptions are shaped as much by social information as by personal experience.

A fourth contribution is to clarify that explanations are not a substitute for well-designed incentives and enforceable safeguards in delegated, data-driven environments. Put simply, even a clear explanation

cannot guarantee good outcomes if the intermediary has reasons to optimize for objectives that differ from the user’s welfare, or if there is no credible way to verify and deter harmful practices. Information disclosure has a long and mixed history as a regulatory tool. In theory, disclosure can improve market outcomes by reducing asymmetric information and enabling market discipline through unraveling and quality sorting (Grossman, 1981; Milgrom, 1981). In practice, however, disclosure regimes often underperform because information is complex, attention is limited, and the content and timing of disclosure can be strategically designed rather than purely informative (Fung et al., 2007; Dranove & Jin, 2010; Y. Bakos et al., 2014).

The dissertation’s evidence fits this broader lesson and helps specify the mechanism. Across the essays, explanations and transparency measures reliably move beliefs and perceived understanding, but their translation into choices is bounded by (i) the payoff structure individuals face, (ii) the interface and attention constraints that govern whether information is processed, and (iii) the contextual nature of trust and acceptability. In economic terms, many of the relevant frictions are not only informational: they are also agency and contracting frictions. In robo-advising and platform recommendation, the decision-maker delegates to an intermediary whose objectives may not coincide with the user’s welfare, creating a principal-agent problem in which effort, objectives, and quality are imperfectly observable (Ross, 1973; Jensen & Meckling, 1976; Holmström, 1979; Grossman & Hart, 1983; Holmström & Milgrom, 1991).

Explanations can mitigate the information problem, but they do not by themselves make the intermediary’s incentives internalize harms, nor do they guarantee that the system is optimized for the outcomes society cares about. This is particularly salient for algorithmic decision-making systems: even when explanations increase perceived understanding or acceptance, they may simply make delegation more comfortable without changing the underlying incentives, error costs, or externalities documented across contexts in the dissertation. The policy implication is therefore to treat explanations as one component of a broader regulatory menu that includes monitoring and enforcement capacity so that quality can be verified and harmful behavior can be deterred rather than merely disclosed (Kroll et al., 2017).

A final contribution is to establish that context dependence is a core empirical fact in data-driven decision-making. Here, “context” is not a label for a sector or a domain; it is the economic decision environment that bundles together (i) what is at stake, (ii) what kind of information is at issue, (iii) what social expectations and norms govern its use, and (iv) which interface mediates the choice. Across the dissertation, the same transparency tool, the same privacy probability, or even the same delegation opportunity can generate very different responses once these contextual elements shift. This emphasis resonates with the privacy literature’s repeated finding that valuation is structured by the *type* and *use* of information, not only by stable “privacy preferences” (Nissenbaum, 2004; Krasnova et al., 2009; Mothersbaugh et al., 2012; Morando et al., 2014). It also aligns with evidence that privacy choices and disclosures

are inherently contextual in data-rich interactive systems (Knijnenburg & Kobsa, 2013). Consistent with this view, the dissertation’s privacy benchmark focuses on financial data and on immediate, realized consequences, thereby isolating a context where risks are unusually transparent and easy to evaluate. The broader point is that shifting the context (data sensitivity, timing, and observability) can mechanically re-rank what individuals treat as salient and worth trading off (Carrascal et al., 2013).

The same logic applies to transparency and delegation. Explanations are not read, interpreted, or trusted in the abstract: whether they matter depends on the conditions under which the decision is made. For example, whether users are mobile, time-constrained, multitasking, or unfamiliar with the service. A growing body of work shows that device environments affect attention, deliberation, and impulsivity, which provides a natural lens on why explanation effects concentrate among subgroups and interface conditions rather than translating into uniform average shifts (R. J.-H. Wang et al., 2015; Lurie et al., 2018; Mograbi, 2022). In parallel, the large-scale ADM evidence in the dissertation shows that the determinants of acceptance vary systematically with the context of use: some settings foreground privacy, others fairness or accountability, and explanations can either increase confidence or make trade-offs more salient (Dietvorst et al., 2015; M. K. Lee, 2018; Logg et al., 2019; Aysolmaz et al., 2023). The implication is that “transparency” should be modeled and evaluated as a context-specific intervention whose welfare effects depend on the surrounding frictions in the choice environment, including information acquisition costs and limited attention (Stigler, 1961; Sims, 2003; Y. Bakos et al., 2014). Finally, this context perspective suggests an extension of the research agenda: future work should build taxonomies of contexts of use (stakes, sensitivity, device, experience) and estimate which contextual features predict when transparency tools meaningfully change behavior, when they mainly shift beliefs, and when they risk legitimizing systems without improving outcomes.

Policy Implications

A first implication concerns privacy policy and argues for moving beyond notice-and-consent. The first essay points toward a sobering conclusion for notice-and-consent models. If privacy risk receives relatively low weight even under explicit probabilities and immediate feedback, then disclosure-based approaches may underperform in real environments characterized by ambiguity, delayed harms, and complex data flows. This reinforces the case for complementing transparency with structural measures: stronger default protections, limits on secondary use, better security standards, and enforcement that changes the expected cost of misuse. At the same time, the documented role of social information suggests that communication policies matter. Breach notification rules, standardized reporting, and public registries can shift collective salience and influence behavior. But relying on salience shocks is not

a stable governance model; it can generate overreactions, fatigue, or unequal impacts across populations with different informational environments.

A second implication concerns transparency requirements and treats them as interventions that should be tested empirically. Across the dissertation, transparency often changes perceived understanding more than it changes behavior. This does not mean transparency is pointless; perceived understanding can itself be welfare-relevant (e.g., autonomy, procedural legitimacy). But it does imply that regulators should be cautious about promising behavioral improvements from transparency alone. A practical implication is that evaluation should focus on the interface level. Explanations that are effective on paper may fail on mobile devices; explanations that help first-time users may not matter for experienced users; and explanations that work for a highly engaged minority may have near-zero average effects. Disclosure regimes have long suffered from these problems (Fung et al., 2007; McDonald & Cranor, 2008; Dranove & Jin, 2010; Y. Bakos et al., 2014). The dissertation suggests that algorithmic transparency is likely to face the same fate unless it is designed around attention constraints and tested empirically.

A third implication concerns governance and emphasizes that accountability and alignment require more than transparency. The dissertation supports the institutional relevance of explanations, but it also motivates a layered governance approach. Explanations can help users understand and potentially contest outcomes, yet they do not ensure that systems are aligned with societal objectives. Accountability mechanisms such as audits, documentation, impact assessments, and channels for redress are therefore complements rather than substitutes (Kroll et al., 2017). In domains where delegation has high stakes, governance must not rely solely on the user’s capacity to read and interpret explanations.

A fourth implication concerns platform policy and highlights interactions among transparency, competition, and distribution. In platform settings, explanations may alter engagement and consumption in ways that have distributional consequences (e.g., for artists in streaming). Explanations can also interact with platform power: if explanation features increase the stickiness of a platform’s curated ecosystem, they may strengthen incumbency advantages. This suggests that transparency should be analyzed jointly with competition policy and with the economics of promotion and visibility (Aguilar & Waldfogel, 2021).

Limitations

Several limitations qualify how the results should be interpreted and generalized. In each case, the point is not that the evidence is fragile, but that every empirical design trades off scope, control, and realism in ways that matter for what we can ultimately claim.

First, the dissertation deliberately combines designs with different strengths in terms of generalization. Essay 1 relies on an incentivized benchmark in which privacy risk is explicit, time-varying, and immedi-

ately realized. This stylization is a feature: it isolates the core trade-off and yields clear causal evidence on how individuals weigh probabilistic privacy harms relative to monetary incentives. The corresponding limitation is that such a benchmark cannot reproduce the full complexity of real data ecosystems, where risks are ambiguous, downstream uses are multi-layered, and consequences may be delayed or difficult to attribute. By contrast, Essay 2 is embedded in an applied robo-advising environment and therefore benefits from strong ecological validity: the decisions studied are close to the way individuals actually encounter algorithmic advice in high-stakes financial settings. Essay 4 similarly improves external validity on the attitudinal margin by relying on a large-scale design that compares responses across multiple contexts and respondent types. The trade-off there is that it measures beliefs, perceptions, and stated willingness to delegate rather than realized behavior in naturally occurring decision processes. Taken together, the dissertation addresses these concerns by combining different methods and settings, but conclusions should still be interpreted in light of the design used in each essay.

Second, the dissertation provides evidence on learning and updating, but it primarily captures short-to medium-run responses rather than long-run dynamics. In Essays 2 and 3, explanations are introduced, and their immediate effects on perceptions, engagement, and choices are measured; in Essay 1, repeated decisions with feedback allow within-experiment updating and reveal the role of social information. The limitation is that transparency and privacy responses may evolve over longer horizons: explanations might become more useful as individuals learn to interpret them, or they might lose salience through habituation; similarly, social learning may operate through slower diffusion processes, media cycles, and network effects that are not fully observed in a controlled setting. More broadly, the dissertation argues that misaligned objectives and delegation frictions can limit what explanations accomplish, but it does not observe how providers adapt their systems over time in response to user behavior or regulation. Addressing these dynamic issues would require longitudinal designs, repeated-interaction settings, and—ideally—platform or institutional partnerships that track behavior, incidents, and belief formation over time.

Third, the dissertation treats explanations as governance-relevant interventions, but it cannot fully rule out strategic design and strategic selection of explanation formats by providers. The empirical chapters study how users react to explanations when they are provided, which is essential for assessing whether transparency can plausibly improve understanding or trust. The limitation is that in real markets, explanations may be optimized for acceptance, reputational protection, or compliance, rather than for comprehension. At the same time, the possibility of strategic explanations does not imply that opacity is preferable: the *absence* of explanations can exacerbate information asymmetries and reduce contestability, leaving users with fewer tools to interpret or challenge algorithmic outputs. The appropriate conclusion is therefore balanced: explanations are valuable as part of the information environment, but their benefits depend on standards, presentation, and the surrounding safeguards that discourage

“performative” compliance.

Finally, heterogeneity raises both interpretive and normative questions that the dissertation highlights but does not fully resolve. The empirical results repeatedly indicate that effects concentrate among subgroups and contexts—depending on attention, device constraints, experience, and baseline attitudes—meaning that policies based on average effects may miss distributional consequences. This links directly to concerns about AI literacy and digital inequality: individuals with lower digital skills, limited access to information, or accessibility barriers may be disproportionately *left out* of the benefits of transparency, and may also be more exposed to its costs. A central open question is therefore how to design transparency and privacy policies that are not only effective on average, but also inclusive—reducing rather than widening informational gaps across the populations most affected by algorithmic systems.

A research agenda

The dissertation suggests a research agenda centered on a simple idea: in data-rich environments, welfare and behavior are shaped not only by what information exists, but by how information interacts with incentives, attention, social learning, and institutional safeguards. This perspective naturally connects the four essays to broader themes in information economics and behavioral economics, where frictions arise from asymmetric information, endogenous information acquisition, and the costs of processing and acting on information (Grossman, 1981; Milgrom, 1981; Sims, 2003; Matějka & McKay, 2015).

A first priority is to move from one-shot or short-horizon evidence toward a deeper understanding of *dynamic* privacy and delegation decisions. Privacy risks are often experienced as delayed, uncertain, and difficult to price, which makes learning and belief formation central (Nissenbaum, 2004; Dinev & Hart, 2006; Acquisti et al., 2016). The privacy benchmark in the dissertation intentionally gives individuals unusually clear probabilities and immediate feedback, thereby creating a favorable environment for a privacy-calculus interpretation; future work should now ask how these trade-offs evolve when probabilities are imprecise, feedback is noisy, and outcomes arrive with long delays. This agenda is naturally related to experimental evidence that individual behavior under risk and ambiguity is heterogeneous and sometimes unstable across elicitation procedures and frames (M. Cohen, Tallon, & Vergnaud, 2011; Hollard, Maafi, & Vergnaud, 2016). It also resonates with experimental approaches that study choices when outcomes are probabilistic and socially meaningful, highlighting how individuals interpret and respond to chance-based consequences (Krawczyk & Le Lec, 2010). Methodologically, this points toward longitudinal designs and repeated-interaction settings that can separate learning and habituation, and can trace how experiences translate into updated beliefs and subsequent delegation or privacy-protective behavior (Sims, 2003; Matějka & McKay, 2015).

A second direction is to treat transparency as an *attention allocation* problem as much as an information-design problem. The dissertation’s results across robo-advising and music streaming are consistent with a two-stage logic: explanations may be provided, yet their effects depend on whether they are noticed, processed, and understood at the moment of choice. This interpretation aligns with rational inattention models in which individuals acquire and process information only when its expected value exceeds its cognitive and time costs (Sims, 2003; Matějka & McKay, 2015). It also echoes a large disclosure literature showing that even legally available information can be under-consumed when it is complex, poorly timed, or costly to parse (Culnan & Armstrong, 1999; Fung et al., 2007; Dranove & Jin, 2010; Y. Bakos et al., 2014). A concrete agenda follows from the dissertation’s emphasis on device constraints, interface placement, and user experience: future work should test explanation formats that explicitly lower processing costs, and should evaluate these formats *in situ* in the digital environments where decisions are actually made (Knijnenburg & Kobsa, 2013; Lurie et al., 2018; Mograbi, 2022). More generally, the relevant research question is not only *what* to explain, but *when* and *under which constraints* explanations become decision-relevant, including in repeated interactions where individuals may learn to use explanations more effectively or, conversely, learn to ignore them.

A third direction concerns the welfare consequences of *adaptive* and *personalized* transparency. If attention and literacy constraints are central, tailoring explanations to user needs and contexts can be attractive, especially in complex financial delegation environments. But personalization is not innocuous: it can introduce unequal treatment, obscure comparability, or become a tool for increasing compliance rather than comprehension. This makes the welfare question explicitly distributional: who benefits from adaptive explanations, who is left out, and how do these tools interact with AI literacy and the digital divide? (Moreau et al., 2022). In finance, the stakes are particularly high because advice and delegation decisions compound over long horizons; this motivates future work that jointly studies transparency, suitability, and household welfare in settings where small differences in risk exposure and fees accumulate over time (Capponi et al., 2022). On the design side, an important task is to develop standards for explanation quality and evaluation protocols that remain robust when explanations adapt to the user and the interface (Arrieta et al., 2020; Beaudouin et al., 2020).

A closely related agenda is to clarify the boundary between *informing* and *persuading*. The dissertation’s evidence is compatible with the idea that explanations can reshape beliefs and perceived understanding without producing uniform behavioral effects, which raises a governance concern: explanation systems may be optimized for acceptance, retention, or platform objectives rather than for user welfare. This connects naturally to the economics of persuasion and information design, where senders strategically choose what information to reveal, how to frame it, and when to disclose it (Crawford & Sobel, 1982; Kamenica & Gentzkow, 2011; Bergemann & Morris, 2019). Translating this to algorithmic

environments suggests a research program that treats explanations as strategic objects embedded in market design and platform incentives (Mestikou et al., 2023). Empirically, this calls for designs that can distinguish comprehension from compliance, measure belief updating versus behavioral steering, and evaluate how different explanation regimes shift trust, adoption, and downstream welfare (Adadi & Berrada, 2018; Arrieta et al., 2020).

A fifth direction is to extend the dissertation’s social-learning insights into a broader study of the *collective* determinants of privacy and delegation. The privacy benchmark indicates that information about others’ experiences can shift behavior more strongly and persistently than one’s own experience, consistent with classic models of herding and informational cascades (Banerjee, 1992; Bikhchandani et al., 1992). Future work should therefore identify the channels through which privacy-relevant information diffuses (peers, workplace networks, social media, news), quantify spillovers beyond directly affected users, and study how narratives shape perceived prevalence and severity of risk (Boyd & Marwick, 2011). This agenda can be connected to experimental and field evidence on peer effects and social interactions in networks, including work showing how information flows and social comparisons can shape effort, behavior, and moral choices (Beugnot, Fortin, Lacroix, & Villeval, 2019; Charroin, Fortin, & Villeval, 2022). A particularly policy-relevant question is how breach notification regimes, transparency reports, and media coverage interact with these social channels: they may change behavior primarily by shifting collective salience rather than by improving private information one user at a time.

Finally, the dissertation motivates an agenda on the *joint governance* of privacy and algorithmic decision-making. Policy debates often separate privacy protection from algorithmic transparency, yet the two objectives are intertwined: meaningful explanations may require revealing aspects of data provenance or model logic, while privacy constraints can limit auditing, contestability, and verification. A central research task is to characterize governance designs that preserve privacy while enabling effective oversight, including audits, impact assessments, and contestability mechanisms (Kroll et al., 2017; Kacianka & Pretschner, 2021). Economically, this agenda can be framed as a problem of institutional design under incomplete contracting and multi-dimensional quality, where disclosure helps only if there exist credible mechanisms to verify claims and deter harmful practices (Grossman, 1981; Milgrom, 1981; Dranove & Jin, 2010). Taken together, these directions suggest that the next generation of research should evaluate transparency not as a presumed remedy, but as a testable intervention whose welfare effects depend on incentives, attention, social diffusion, and enforceable safeguards. This evaluation should rely on designs that can credibly scale from controlled settings to high-stakes, real-world environments (Kawaguchi, 2021; Cohodes, 2022; List, 2022).

Bibliography

References

- Abraham, F., Schumaker, S. L., & Tessada, J. (2019). *Robo-advisors : Investing through machines* (Research and Policy Briefs No. 134881). The World Bank.
- Acquisti, A. (2014). The economics and behavioral economics of privacy. *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, 1, 76–95.
- Acquisti, A., & Gross, R. (2006). Imagined communities: Awareness, information sharing, and privacy on the facebook. In *International workshop on privacy enhancing technologies* (pp. 36–58).
- Acquisti, A., & Grossklags, J. (2005). Privacy and rationality in individual decision making. *IEEE Security & Privacy*, 3(1), 26–33.
- Acquisti, A., Grossklags, J., & et al. (2007). What can behavioral economics teach us about privacy. *Digital Privacy: Theory, Technologies and Practices*, 18, 363–377.
- Acquisti, A., Taylor, C., & Wagman, L. (2016). The economics of privacy. *Journal of Economic Literature*, 54(2), 442–492. doi: 10.1257/jel.54.2.442
- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE Access*, 6, 52138–52160.
- Adomavicius, G., Bockstedt, J. C., Curley, S. P., & Zhang, J. (2018). Effects of online recommendations on consumers’ willingness to pay. *Information Systems Research*, 29(1), 84–102.
- Afchar, D., Melchiorre, A., Schedl, M., Hennequin, R., Epure, E., & Moussallam, M. (2022). Explainability in music recommender systems. *AI Magazine*, 43(2), 190–208.
- Agnew, J., & Szykman, L. (2005). Asset allocation and information overload: The influence of information display, asset choice, and investor experience. *Journal of Behavioral Finance*, 6(2), 57–70. doi: 10.1207/s15427579jpfm0602_2
- Aguiar, L., & Waldfogel, J. (2021). Platforms, power, and promotion: Evidence from spotify playlists. *The Journal of Industrial Economics*, 69(3), 653–691.
- Akerlof, G. A. (1970). The market for “lemons”: Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3), 488–500. doi: 10.2307/1879431
- Altman, I. (1975). The environment and social behavior: privacy, personal space, territory, and crowding.
- Andjelkovic, I., Parra, D., & O’Donovan, J. (2016). Moodplay: Interactive mood-based music discovery and recommendation. In *Proceedings of the 2016 conference on user modeling adaptation and personalization* (pp. 275–279).
- Angel, S., Disslbacher, F., Humer, S., & Schnetzer, M. (2019). What did you really earn last year?: explaining measurement error in survey income data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 182(4), 1411–1437.

- Anjomshoae, S., Najjar, A., Calvaresi, D., & Främling, K. (2019). Explainable agents and robots: Results from a systematic literature review. In *18th international conference on autonomous agents and multiagent systems (aamas 2019), montreal, canada, may 13–17, 2019* (pp. 1078–1088).
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58, 82–115.
- Aysolmaz, B., Müller, R., & Meacham, D. (2023). The public perceptions of algorithmic decision-making systems: Results from a large-scale survey. *Telematics and Informatics*, 79, 101954.
- Bakos, J. Y. (1997). Reducing buyer search costs: Implications for electronic marketplaces. *Management Science*, 43(12), 1676–1692.
- Bakos, Y., Marotta-Wurgler, F., & Trossen, D. R. (2014). Does anyone read the fine print? consumer attention to standard-form contracts. *The Journal of Legal Studies*, 43(1), 1–35.
- Banerjee, A. V. (1992). A simple model of herd behavior. *The Quarterly Journal of Economics*, 107(3), 797–817. doi: 10.2307/2118364
- Bansal, G., Zahedi, F. & Gefen, D. (2015). The role of privacy assurance mechanisms in building trust and the moderating role of privacy concern. *European Journal of Information Systems*, 24(6), 624–644.
- Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6), 1173–1182. doi: 10.1037/0022-3514.51.6.1173
- Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2018). Consumer-lending discrimination in the fintech era. *Journal of Financial Economics*, 143(1), 30–56. doi: 10.1016/j.jfineco.2021.05.047
- Bartling, B., Fehr, E., & Herz, H. (2014). The intrinsic value of decision rights. *Econometrica*, 82(6), 2005–2039.
- Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods*, 11(2), 142–163. doi: 10.1037/1082-989X.11.2.142
- Beaudouin, V., Bloch, I., Bounie, D., Cléménçon, S., d’Alché Buc, F., Eagan, J., ... Parekh, J. (2020). Flexible and context-specific ai explainability: a multidisciplinary approach. *arXiv preprint arXiv:2003.07703*.
- Benartzi, S., & Thaler, R. (2001). Naive diversification strategies in defined contribution saving plans. *American Economic Review*, 91, 79–98. doi: 10.1257/aer.91.1.79
- Ben David, D., Resheff, Y. S., & Tron, T. (2021). Explainable ai and adoption of financial algorithmic advisors: an experimental study. In *Proceedings of the 2021 aaai/acm conference on ai, ethics, and*

- society* (pp. 390–400).
- Ben-Shahar, O., & Schneider, C. E. (2014). *More than you wanted to know: The failure of mandated disclosure*. Princeton, NJ: Princeton University Press. Retrieved from <https://academic.oup.com/princeton-scholarship-online/book/18199>
- Berendt, B., Günther, O., & Spiekermann, S. (2005). Privacy in e-commerce: Stated preferences vs. actual behavior. *Communications of the ACM*, 48(4), 101–106.
- Bergemann, D., & Morris, S. (2019). Information design: A unified perspective. *Journal of Economic Literature*, 57(1), 44–95.
- Beshears, J., & et al. (2020). Nudging: Progress to date and future directions. *Behavioural Public Policy*.
- Beugnot, J., Fortin, B., Lacroix, G., & Villeval, M. C. (2019). Gender and peer effects on performance in social networks. *European Economic Review*, 113, 207–224. doi: 10.1016/j.euroecorev.2019.01.002
- Bianchi, M., & Brière, M. (2020). Robo-advising for small investors. *SSRN Electronic Journal*.
- Bianchi, M., & Briere, M. (2021). Robo-advising: less ai and more xai? *Available at SSRN 3825110*.
- Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5), 992–1026. doi: 10.1086/261849
- Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). 'it's reducing a human being to a percentage' perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1–14).
- Bloom, N., Sadun, R., & Van Reenen, J. (2012). The organization of firms across countries. *The Quarterly Journal of Economics*, 127(4), 1663–1705.
- Boissay, F., Ehlers, T., Gambacorta, L., & Shin, H. S. (2021). Big techs in finance: on the new nexus between data privacy and competition. In *The palgrave handbook of technological finance* (pp. 855–875). Springer.
- Boulu-Reshef, B., & Monnier, C. (2024). Punitive attitudes toward cartels: evidence from an experimental study. *European Journal of Law and Economics*, 58(3), 481–516.
- Bound, J., Brown, C., & Mathiowetz, N. (2001). Measurement error in survey data. In *Handbook of econometrics* (Vol. 5, pp. 3705–3843). Elsevier.
- Boyd, D., & Marwick, A. E. (2011). Social privacy in networked publics: Teens' attitudes, practices, and strategies. In *A decade in internet time: Symposium on the dynamics of the internet and society*.
- Cai, C. J., Jongejan, J., & Holbrook, J. (2019). The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces* (pp. 258–262).
- Capponi, A., Olafsson, S., & Zariphopoulou, T. (2022). Personalized robo-advising: Enhancing invest-

- ment through client interaction. *Management Science*, 68(4), 2485–2512.
- Cardillo, G., & Chiappini, H. (2024). Robo-advisors: A systematic literature review. *Finance Research Letters*, 62, 105119. Retrieved from <https://www.sciencedirect.com/science/article/pii/S1544612324001491> doi: <https://doi.org/10.1016/j.frl.2024.105119>
- Carrascal, J. P., Riederer, C., Erramilli, V., Cherubini, M., & De Oliveira, R. (2013). Your browsing behavior for a big mac: Economics of personal information online. In *Proceedings of the 22nd international conference on world wide web* (pp. 189–200).
- Castelluccia, C., & Le Métayer, D. (2019). *Understanding algorithmic decision-making – opportunities and challenges*. Publications Office. doi: [doi/10.2861/536131](https://doi.org/10.2861/536131)
- Castelvecchi, D. (2016). Can we open the black box of ai? *Nature*, 538(7623), 20–23.
- Charness, G., & Gneezy, U. (2009). Incentives to exercise. *Econometrica*, 77(3), 909–931.
- Charroin, L., Fortin, B., & Villeval, M. C. (2022). Peer effects, self-selection and dishonesty. *Journal of Economic Behavior & Organization*, 200, 618–637. doi: [10.1016/j.jebo.2022.06.024](https://doi.org/10.1016/j.jebo.2022.06.024)
- Chen, D. L., Schonger, M., & Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9, 88–97.
- Chen, J., Song, L., Wainwright, M., & Jordan, M. (2018). Learning to explain: An information-theoretic perspective on model interpretation. In *International conference on machine learning* (pp. 883–892).
- Cheng, H.-F., Wang, R., Zhang, Z., O’connell, F., Gray, T., Harper, F. M., & Zhu, H. (2019). Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1–12).
- Chetty, R., Looney, A., & Kroft, K. (2009). Salience and taxation: Theory and evidence. *American Economic Review*, 99(4), 1145–1177. doi: [10.1257/aer.99.4.1145](https://doi.org/10.1257/aer.99.4.1145)
- Cingano, F., & Pinotti, P. (2012). Trust, firm organization and the structure of production. *Paolo Baffi Centre Research Paper*(2012-133).
- Cirqueira, D., Nedbal, D., Helfert, M., & Bezbradica, M. (2020). Scenario-based requirements elicitation for user-centric explainable ai: A case in fraud detection. In *International cross-domain conference for machine learning and knowledge extraction* (pp. 321–341).
- Cohen, M., Tallon, J.-M., & Vergnaud, J.-C. (2011). An experimental investigation of imprecision attitude and its relation with risk attitude and impatience. *Theory and Decision*, 71, 81–109. doi: [10.1007/s11238-010-9205-8](https://doi.org/10.1007/s11238-010-9205-8)
- Cohen, S. B., Dror, G., & Ruppin, E. (2005). Feature selection based on the shapley value. In *Proceedings of ijcai* (pp. 1–6).
- Cohodes, S. R. (2022). *When do informational interventions work? experimental evidence from new york*

- city high school choice* (Tech. Rep. No. 29690). NBER.
- Cooper, D. J., Ioannou, C. A., & Qi, S. (2018). Endogenous incentive contracts and efficient coordination. *Games and Economic Behavior*, 112, 78–97.
- Covert, I., Lundberg, S., & Lee, S.-I. (2021). Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209), 1–90.
- Cramer, H., Evers, V., Ramlal, S., Van Someren, M., Rutledge, L., Stash, N., ... Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-adapted interaction*, 18(5), 455–496.
- Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. *Econometrica*, 50(6), 1431–1451.
- Culnan, M. J., & Armstrong, P. K. (1999). Information privacy concerns, procedural fairness, and impersonal trust: An empirical investigation. *Organization Science*, 10(1), 104–115.
- D’Acunto, F., & Rossi, A. G. (2023). Robo-advice: Transforming households into rational economic agents. *Annual Review of Financial Economics*, 15(1), 543–563.
- d’Astous, P., Gemmo, I., & Michaud, P.-C. (2022). *The quality of financial advice: What influences client recommendations?* (Tech. Rep.). National Bureau of Economic Research.
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319–340.
- De Bruyn, A., Viswanathan, V., Beh, Y. S., Brock, J. K.-U., & Von Wangenheim, F. (2020). Artificial intelligence and marketing: Pitfalls and opportunities. *Journal of Interactive Marketing*, 51(1), 91–105.
- DellaVigna, S. (2009). Psychology and economics: Evidence from the field. *Journal of Economic Literature*, 47(2), 315–372.
- DellaVigna, S., & Gentzkow, M. (2010). Persuasion: Empirical evidence. *Annual Review of Economics*, 2(1), 643–669.
- Deo, S., & Sontakke, N. S. (2021). Usability, user comprehension, and perceptions of explanations for complex decision support systems in finance: A robo-advisory use case. *Computer*, 54(10), 38–48.
- Diab, D. L., Pui, S.-Y., Yankelevich, M., & Highhouse, S. (2011). Lay perceptions of selection decision aids in us and non-us samples. *International Journal of Selection and Assessment*, 19(2), 209–216.
- Dienlin, T., & Trepte, S. (2015). Is the privacy paradox a relic of the past? an in-depth analysis of privacy attitudes and privacy behaviors. *European Journal of Social Psychology*, 45(3), 285–297.
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General*, 144(1), 114.
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management science*, 64(3),

1155–1170.

- Dinev, T., & Hart, P. (2006). An extended privacy calculus model for e-commerce transactions. *Information Systems Research*, 17(1), 61–80.
- Dominitz, J., & Hung, A. (2006). Retirement savings portfolio management. *SSRN Electronic Journal*, 138. doi: 10.2139/ssrn.1095290
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- Dranove, D., & Jin, G. Z. (2010). Quality disclosure and certification: Theory and practice. *Journal of Economic Literature*, 48(4), 935–963. doi: 10.1257/jel.48.4.935
- Dumitrescu, E. I., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, 297(3), 1178–1192. doi: 10.1016/j.ejor.2021.06.053
- Eslami, M., Krishna Kumaran, S. R., Sandvig, C., & Karahalios, K. (2018). Communicating algorithmic process in online behavioral advertising. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1–13).
- Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., ... Sandvig, C. (2015). “I always assumed that I wasn’t really that close to [her]”: Reasoning about invisible algorithms in news feeds. In *Proceedings of the 33rd annual acm conference on human factors in computing systems* (pp. 153–162). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2702123.2702556
- European Commission. (2019). *Ethics guidelines for trustworthy AI*. Retrieved from <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- European Parliament, & Council of the European Union. (2024). *Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act)*. Official Journal of the European Union, L_202401689, 12 July 2024. Retrieved from <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng> (CELEX: 32024R1689)
- Farchy, J., Méadel, C., & Anciaux, A. (2017). Une question de comportement. recommandation des contenus audiovisuels et transformations numériques. *Tic&société*, 10(2-3), 168–198.
- Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4), 980–994.
- Fein, M. L. (2017). Are robo-advisors fiduciaries? *Available at SSRN 3028268*.
- Foerster, S., Linnainmaa, J., Melzer, B., & Previtero, A. (2018). Financial advisors and risk-taking. In *Assa annual meeting, association of financial economists*. Philadelphia: American Economic

- Association.
- Friedman, D., & Sunder, S. (1994). *Experimental methods: A primer for economists*. Cambridge University Press.
- Frik, A., & Mittone, L. (2019). Factors influencing the perception of website privacy trustworthiness and users' purchasing intentions: The behavioral economics perspective. *Journal of Theoretical and Applied Electronic Commerce Research*, 14(3), 89–125.
- Fung, A., Graham, M., & Weil, D. (2007). *Full disclosure: The perils and promise of transparency*. Cambridge: Cambridge University Press.
- Gabaix, X. (2019). Behavioral inattention. In B. D. Bernheim, S. DellaVigna, & D. Laibson (Eds.), *Handbook of behavioral economics: Applications and foundations 1* (Vol. 2, pp. 261–343). Elsevier. doi: 10.1016/bs.hesbe.2018.11.001
- Gabaix, X., & Laibson, D. (2006). Shrouded attributes, consumer myopia, and information suppression in competitive markets. *The Quarterly Journal of Economics*, 121(2), 505–540. doi: 10.1162/qjec.2006.121.2.505
- Gächter, S., & Herrmann, B. (2009). Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1518), 791–806.
- Gächter, S., Johnson, E. J., & Herrmann, A. (2022). Individual-level loss aversion in riskless and risky choices. *Theory and Decision*, 92(3-4), 599–624.
- Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and tam in online shopping: An integrated model. *MIS Quarterly*, 51–90.
- Gennaioli, N., Shleifer, A., & Vishny, R. (2015). Money doctors. *The Journal of Finance*, 70(1), 91–114.
- Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, 62(2011), 451–482.
- Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*, 103(4), 650.
- Gneezy, U., & Potters, J. (1997). An experiment on risk taking and evaluation periods. *The Quarterly Journal of Economics*, 112(2), 631–645.
- Goldfarb, A., & Tucker, C. (2012). Shifts in privacy concerns. *American Economic Review*, 102(3), 349–353.
- Grossman, S. J. (1981). The informational role of warranties and private disclosure about product quality. *Journal of Law and Economics*, 24(3), 461–483. doi: 10.1086/466999
- Grossman, S. J., & Hart, O. D. (1983). An analysis of the principal-agent problem. *Econometrica*, 51(1), 7–45. doi: 10.2307/1912246

- Grossman, S. J., & Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *The American Economic Review*, 70(3), 393–408.
- Guiso, L., Jappelli, T., & Terlizzese, D. (1996). *Income risk, borrowing constraints, and portfolio choice* (Vol. 86) (No. 1). American Economic Association.
- Guiso, L., Sapienza, P., & Zingales, L. (2008). Trusting the stock market. *the Journal of Finance*, 63(6), 2557–2600.
- Gur, N., & Bjørnskov, C. (2017). Trust and delegation: Theory and evidence. *Journal of Comparative Economics*, 45(3), 644–657.
- Haque, A., Pranto, T. H., Noman, A. A., & Mahmood, A. (2021). Insight about detection, prediction and weather impact of coronavirus (covid-19) using neural network. *arXiv preprint arXiv:2104.02173*.
- Heidhues, P., Johnen, J., & Kőszegi, B. (2021). Browsing versus studying: A pro-market case for regulation. *The Review of Economic Studies*, 88(2), 708–729.
- Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). Explaining collaborative filtering recommendations. In *Proceedings of the 2000 acm conference on computer-supported cooperative work* (pp. 241–250).
- Herrmann, B., Thoni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, 319(5868), 1362–1367.
- Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist*, 41(2), 111–127.
- Hollard, G., Maafi, H., & Vergnaud, J. C. (2016). Consistent inconsistencies? evidence from decision under risk. *Theory and Decision*, 80(4), 623–648. doi: 10.1007/s11238-015-9518-8
- Holmström, B. (1979). Moral hazard and observability. *The Bell Journal of Economics*, 10(1), 74–91. doi: 10.2307/3003320
- Holmström, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization*, 7(Special Issue), 24–52.
- Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, 92(5), 1644–1655.
- Holzmeister, F., Holmén, M., Kirchler, M., Stefan, M., & Wengström, E. (2023). Delegation decisions in finance. *Management Science*, 69(8), 4828–4844.
- Hong, C. Y., Lu, X., & Pan, J. (2020). *Fintech adoption and household risk-taking: From digital payments to platform investments* (Tech. Rep.). National Bureau of Economic Research.
- Hu, D., Guo, F., Shang, J., & Zhang, X. (2024). Does digital finance increase household risk-taking? evidence from china. *International Review of Economics & Finance*, 93, 1197–1210.
- Huang, M.-H., & Rust, R. T. (2021). A strategic framework for artificial intelligence in marketing. *Journal of the Academy of Marketing Science*, 49, 30–50.

- Huberman, G. (2001). Familiarity breeds investment. *The Review of Financial Studies*, 14(3), 659–680. doi: 10.1093/rfs/14.3.659
- Hu  , S., Hurlin, C., P  rignon, C., & Saurin, S. (2025). Measuring the driving forces of predictive performance: Application to credit scoring. *HEC Paris Research Paper No. FIN-2022-1463*.
- Hung, A., & Yoong, J. (2013). Asking for help: Survey and experimental evidence on financial advice and behavior change. In *The market for retirement financial advice* (p. 182-212). Oxford University Press. doi: 10.1093/acprof:oso/9780199683772.003.0009
- Hurlin, C., & P  rignon, C. (2019). Machine learning et nouvelles sources de donn  es pour le scoring de cr  dit. *Revue d'  conomie Financ  re*, 135(3), 21–50. doi: 10.3917/ecofi.135.0021
- Hyman, J. M. (2020). Can light-touch college-going interventions make a difference? evidence from a statewide experiment in michigan. *Journal of Policy Analysis and Management*, 39(1), 159–190. doi: 10.1002/pam.22155
- Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15(4), 309–334. doi: 10.1037/a0020761
- Jacob, D. (2021). Cate meets ml: Conditional average treatment effect and machine learning. *Digital Finance*, 3(2), 99–148.
- Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, 3(4), 305–360. doi: 10.1016/0304-405X(76)90026-X
- Jentzsch, N., Preibusch, S., & Harasser, A. (2012). Study on monetising privacy: An economic model for pricing personal information. *ENISA*, 1(1).
- Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1), 53–71.
- Jiang, Z., Heng, C. S., & Choi, B. C. (2013). Research note—privacy concerns and privacy-protective behavior in synchronous online social interactions. *Information Systems Research*, 24(3), 579–595.
- Jung, D., Dorner, V., Weinhardt, C., & Puzmaz, H. (2018). Designing a robo-advisor for risk-averse, low-budget consumers. *Electronic Markets*, 28(3), 367-380. doi: 10.1007/s12525-017-0279-9
- Kacianka, S., & Pretschner, A. (2021). Designing accountable systems. In *Proceedings of the 2021 acm conference on fairness, accountability, and transparency* (pp. 424–437).
- Kahn, C. M., & Roberds, W. (2008). Credit and identity theft. *Journal of Monetary Economics*, 55(2), 251–264.
- Kahneman, D. (1973). *Attention and effort*. Prentice-Hall.
- Kahneman, D. (2013). A perspective on judgment and choice: Mapping bounded rationality. *Progress in Psychological Science around the World. Volume 1 Neural, Cognitive and Developmental Issues.*,

1–47.

- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–291. doi: 10.2307/1914185
- Kamenica, E., & Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, 101(6), 2590–2615.
- Kang, R., Brown, S., & Kiesler, S. (2013). Why do people seek anonymity on the internet? informing policy and design. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 2657–2666).
- Kawaguchi, K. (2021). When will workers follow an algorithm? a field experiment with a retail business. *Management Science*, 67(3), 1670–1695.
- Kizilcec, R. F. (2016). How much information?: Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 chi conference on human factors in computing systems* (pp. 2390–2395). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2858036.2858402
- Knight, F. H. (1921). *Risk, uncertainty and profit* (Vol. 31). Houghton Mifflin.
- Knijnenburg, B. P., & Kobsa, A. (2013). Making decisions about privacy: information disclosure in context-aware recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(3), 1–23.
- Krasnova, H., Günther, O., Spiekermann, S., & Koroleva, K. (2009). Privacy concerns and identity in online social networks. *Identity in the Information Society*, 2, 39–63.
- Krawczyk, M., & Le Lec, F. (2010). ‘Give me a chance!’ an experiment in social decision under risk. *Experimental Economics*, 13(4), 500–511. doi: 10.1007/s10683-010-9253-5
- Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, 165(3), 633–705.
- Krzyżiński, M., Spytek, M., Baniecki, H., & Biecek, P. (2023). Survshap (t): time-dependent explanations of machine learning survival models. *Knowledge-Based Systems*, 262, 110234.
- Laufer, R. S., & Wolfe, M. (1977). Privacy as a concept and a social issue: A multidimensional developmental theory. *Journal of Social Issues*, 33(3), 22–42.
- Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big data & Society*, 5(1), 2053951718756684.
- Lee, N., & Kwon, O. (2015). A privacy-aware feature selection method for solving the personalization–privacy paradox in mobile wellness healthcare services. *Expert Systems with Applications*, 42(5), 2764–2771.
- Linnainmaa, J., Melzer, B., Previtero, A., & Foerster, S. (2020). *Investor protections and stock market*

- participation: an evaluation of financial advisor oversight* (Tech. Rep.). Working paper.
- List, J. A. (2022). *The voltage effect: How to make good ideas great and great ideas scale*. Crown Currency.
- List, J. A. (2023). The voltage effect. *Revue Économique*, 74(5), 673–680.
- Liu, X., Zhou, Y., & Wang, Z. (2019). Recognition and extraction of named entities in online medical diagnosis data based on a deep neural network. *Journal of Visual Communication and Image Representation*, 60, 1–15.
- Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151, 90–103. doi: 10.1016/j.obhdp.2018.12.005
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Lurie, N. H., Berger, J., Chen, Z., Li, B., Liu, H., Mason, C. H., . . . et al. (2018). Everywhere and at all times: mobility, consumer decision-making, and choice. *Customer Needs and Solutions*, 5, 15–27.
- Mann, C. L. (2015). Information lost: Will the "paradise" that information promises, to both consumer and firm, be "lost" on account of data breaches? the epic is playing out. *Economic Analysis of the Digital Economy*, 309–351.
- Martijn, M., Conati, C., & Verbert, K. (2022). “knowing me, knowing you”: personalized explanations for a music recommender system. *User Modeling and User-Adapted Interaction*, 32(1-2), 215–252.
- Matějka, F., & McKay, A. (2015). Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1), 272–298. doi: 10.1257/aer.20130047
- McDonald, A. M., & Cranor, L. F. (2008). The cost of reading privacy policies. *Isjlp*, 4, 543.
- Mestikou, M. A., Smeti, K. E., & Hachaïchi, Y. (2023). Artificial intelligence and machine learning in financial services: Market developments and financial stability implications. *Financial Stability Board*, 1, 1–6.
- Mikhed, V., & Vogan, M. (2018). How data breaches affect consumer credit. *Journal of Banking & Finance*, 88, 192–207.
- Milgrom, P. R. (1981). Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics*, 12(2), 380–391. doi: 10.2307/3003562
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38. doi: 10.1016/j.artint.2018.07.007
- Mograbi, E. (2022). Decision-makers are more impulsive on smartphones than on computers. *Journal of Behavioral and Experimental Economics*, 100, 101916.
- Moore, J. C., & Welniak, E. J. (2000). Income measurement error in surveys: A review. *Journal of*

Official Statistics, 16(4), 331.

- Morana, S., Gnewuch, U., Jung, D., & Granig, C. (2020). The effect of anthropomorphism on investment decision-making with robo-advisor chatbots. In *Ecis*.
- Morando, F., Iemma, R., & Raiteri, E. (2014). Privacy evaluation: what empirical research on users' valuation of personal data tells us. *Internet Policy Review*, 3(2), 1–12.
- Moreau, F., Haampland, O., Johannessen, R., & Wikstrom, P. (2022). Fairness and royalty payment systems on music streaming platforms. *Available at SSRN 4248175*.
- Moreau, F., & Wikström, P. (2024). *The Impact of Algorithmic and Human Recommendations on Platform User Satisfaction*. Retrieved from <https://hal.science/hal-04553197>
- Mothersbaugh, D. L., Foxx, W. K., Beatty, S. E., & Wang, S. (2012). Disclosure antecedents in an online service context: The role of sensitivity of information. *Journal of Service Research*, 15(1), 76-98. doi: 10.1177/1094670511424924
- Nissenbaum, H. (2004). Privacy as contextual integrity. *Wash. L. Rev.*, 79, 119.
- Norberg, P. A., Horne, D. R., & Horne, D. A. (2007). The privacy paradox: Personal information disclosure intentions versus behaviors. *Journal of Consumer Affairs*, 41(1), 100–126.
- Nunes, I., & Jannach, D. (2017). A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, 27, 393–444.
- OECD. (2023). *Data portability in open banking: Privacy and other cross-cutting issues* (OECD Digital Economy Papers No. 348). OECD Publishing. Retrieved from <https://ideas.repec.org/p/oec/stiaab/348-en.html> doi: 10.1787/6c872949-en
- Özer, Ö., Subramanian, U., & Wang, Y. (2018). Information sharing, advice provision, or delegation: What leads to higher trust and trustworthiness? *Management Science*, 64(1), 474–493.
- Patel, K., & Lincoln, M. (2019). It's not magic: Weighing the risks of ai in financial services. *Report available at <https://www.european-microfinance.org/publication/its-not-magic-weighing-risks-ai-financial-services>*.
- Philippon, T. (2019). On fintech and financial inclusion. *National Bureau of Economic Research*. doi: 10.3386/w26330
- Posner, R. A. (1978). Economic theory of privacy. *Regulation*, 2, 19.
- Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J., & Wallach, H. (2021). Manipulating and measuring model interpretability. In *Proceedings of the 2021 chi conference on human factors in computing systems* (pp. 1–52). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3411764.3445315
- Preece, A., Harborne, D., Braines, D., Tomsett, R., & Chakraborty, S. (2018). Stakeholders in explainable ai. *arXiv preprint arXiv:1810.00184*.

- Preibusch, S., Kübler, D., & Beresford, A. R. (2013). Price versus privacy: an experiment into the competitive advantage of collecting less personal information. *Electronic Commerce Research*, 13, 423–455.
- Rader, E., Cotter, K., & Cho, J. (2018). Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1–13).
- Rainie, L., Kiesler, S., Kang, R., Madden, M., Duggan, M., Brown, S., & Dabbish, L. (2013). Anonymity, privacy, and security online. *Pew Research Center*, 5.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). “why should i trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1135–1144). ACM. doi: 10.1145/2939672.2939778
- Riedl, M. O. (2019). Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, 1(1), 33–36.
- Roberds, W., & Schreft, S. L. (2009). Data breaches and identity theft. *Journal of Monetary Economics*, 56(7), 918–929.
- Rochet, J.-C., & Tirole, J. (2003). Platform competition in two-sided markets. *Journal of the European Economic Association*, 1(4), 990–1029.
- Rochet, J.-C., & Tirole, J. (2006). Two-sided markets: A progress report. *RAND Journal of Economics*, 37(3), 645–667.
- Ross, S. A. (1973). The economic theory of agency: The principal’s problem. *The American Economic Review*, 63(2), 134–139.
- Schoeman, F. D. (1992). *Privacy and social freedom*. Cambridge University Press.
- Schröder, A. M., & Ghajargar, M. (2021). Unboxing the algorithm: Designing an understandable algorithmic experience in music recommender systems. In *15th acm conference on recommender systems (recsys 2021), amsterdam, the netherlands, september 25, 2021*.
- Shen, Y., Hu, W., & Zhang, Y. (2022). Digital finance, household income and household risky financial asset investment. *Procedia Computer Science*, 202, 244–251.
- Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International Journal of Human-Computer Studies*, 146, 102551.
- Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1), 99–118. doi: 10.2307/1884852
- Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, 50(3), 665–690.

doi: 10.1016/S0304-3932(03)00029-1

- Smith, H. J., Milberg, S. J., & Burke, S. J. (1996). Information privacy: Measuring individuals' concerns about organizational practices. *MIS quarterly*, 167–196.
- Smith, V. L. (1976). Experimental economics: Induced value theory. *The American Economic Review*, 66(2), 274–279.
- Solove, D. J. (2005). A taxonomy of privacy. *U. Pa. l. Rev.*, 154, 477.
- Speith, T. (2022). A review of taxonomies of explainable artificial intelligence (xai) methods. In *Proceedings of the 2022 acm conference on fairness, accountability, and transparency* (pp. 2239–2250).
- Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, 87(3), 355–374. doi: 10.2307/1882010
- Srnicek, N. (2017). *Platform capitalism*. Cambridge: Polity.
- Stigler, G. J. (1961). The economics of information. *Journal of Political Economy*, 69(3), 213–225.
- Stiglitz, J. E. (1975). The theory of “screening”, education, and the distribution of income. *The American Economic Review*, 65(3), 283–300.
- Strzelczyk, B. E. (2017). Rise of the machines: The legal implications for investor protection with the rise of robo-advisors. *DePaul Bus. & Comm. LJ*, 16, 54.
- Stutzman, F. D., Gross, R., & Acquisti, A. (2013). Silent listeners: The evolution of privacy and disclosure on facebook. *Journal of Privacy and Confidentiality*, 4(2), 2.
- Sutanto, J., Palme, E., Tan, C.-H., & Phang, C. W. (2013). Addressing the personalization-privacy paradox: An empirical assessment from a field experiment on smartphone users. *MIS Quarterly*, 1141–1164.
- Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. New Haven, CT: Yale University Press.
- Tintarev, N., & Masthoff, J. (2015). Explaining recommendations: Design and evaluation. In *Recommender systems handbook* (pp. 353–382). Springer.
- Tsai, J. Y., Egelman, S., Cranor, L., & Acquisti, A. (2011). The effect of online privacy information on purchasing behavior: An experimental study. *Information Systems Research*, 22(2), 254–268.
- Turow, J., King, J., Hoofnagle, C. J., Bleakley, A., & Hennessy, M. (2009). Americans reject tailored advertising and three activities that enable it. *Available at SSRN 1478214*.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131. doi: 10.1126/science.185.4157.1124
- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3), 425–478.
- Wakefield, R. (2013). The influence of user affect in online information disclosure. *The Journal of*

- Strategic Information Systems*, 22(2), 157–174.
- Wang, R. J.-H., Malthouse, E. C., & Krishnamurthi, L. (2015). On the go: How mobile shopping affects customer purchase behavior. *Journal of Retailing*, 91(2), 217–234.
- Wang, W., & Benbasat, I. (2016). Empirical assessment of alternative designs for enhancing different types of trusting beliefs in online recommendation agents. *Journal of Management Information Systems*, 33(3), 744–775.
- Warren, S., & Brandeis, L. (1989). The right to privacy. In *Killing the messenger: 100 years of media criticism* (pp. 1–21). Columbia University Press.
- Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2021). Let me explain!: exploring the potential of virtual agents in explainable ai interaction design. *Journal on Multimodal User Interfaces*, 15(2), 87–98.
- Westin, A. F. (1968). Privacy and freedom. *Washington and Lee Law Review*, 25(1), 166.
- Wieringa, M. (2020). What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 1–18).
- Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20(7), 557–585.
- Xu, H., Luo, X. R., Carroll, J. M., & Rosson, M. B. (2011). The personalization privacy paradox: An exploratory study of decision making process for location-aware marketing. *Decision Support Systems*, 51(1), 42–52.
- Završnik, A. (2020). Criminal justice, artificial intelligence systems, and human rights. In *Era forum* (Vol. 20, pp. 567–583).
- Zhang, L., Pentina, I., & Fan, Y. (2021). Who do you choose? comparing perceptions of human vs robo-advisor in the context of financial services. *Journal of Services Marketing*, 35(5), 634–646.
- Zhang, Y., & Chen, X. (2020). Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval*, 14(1), 1–101.
- Zhao, G., Fu, H., Song, R., Sakai, T., Chen, Z., Xie, X., & Qian, X. (2019). Personalized reason generation for explainable song recommendation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(4), 1–21.
- Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. New York: PublicAffairs.

Résumé in Extenso

Résumé in Extenso

La prise de décision à l'ère du capitalisme numérique

Une part croissante de la vie économique se déroule désormais dans des environnements numériques : nous investissons par l'intermédiaire de robots-conseillers, découvrons de la musique grâce à des systèmes de recommandation algorithmique, effectuons des achats à partir de classements et interagissons avec des services publics et privés via des interfaces qui prédisent et préremplissent nos préférences. Ces environnements ne se contentent pas d'afficher des informations ; ils façonnent activement ce qui est vu, dans quel ordre et comment cela est justifié. Les systèmes de classement et de personnalisation déterminent ce qui entre dans le champ de réflexion du décideur, tandis que la collecte de données transforme chaque clic, chaque défilement et chaque choix en informations pouvant être utilisées pour prédire et parfois influencer les comportements futurs. Cette thèse étudie la prise de décision dans ce contexte en se concentrant sur l'*Homo Numericus* : des individus dont les choix économiques sont de plus en plus influencés par l'environnement numérique et les systèmes algorithmiques.

Le principe central est qu'une fois que les algorithmes recommandent, classent et personnalisent, la structure d'information pertinente à laquelle sont confrontés les individus change. L'économie classique de l'information souligne que les incitations et les résultats dépendent essentiellement de ce que les agents savent, de ce qu'ils croient et de la manière dont l'information est distribuée et révélée (Stigler, 1961; Akerlof, 1970; Spence, 1973; Stiglitz, 1975; Grossman & Stiglitz, 1980). L'intermédiation numérique modifie cette structure dans deux directions à la fois. Premièrement, l'algorithme devient un producteur d'informations pour l'utilisateur (scores, classements, recommandations). Ensuite, la plateforme devient un extracteur d'informations provenant de l'utilisateur (traces de données, signaux comportementaux), transformant les informations personnelles en données d'entrée et en actifs (Acquisti et al., 2016; Srnicek, 2017; Zuboff, 2019). En conséquence, l'économie de la prise de décision dépend de plus en plus de la manière dont les gens comprennent, font confiance et réagissent aux systèmes qui les conseillent et apprennent d'eux simultanément.

Cette transformation est particulièrement visible sur les marchés *médiés par des plateformes*. Les

plateformes mettent souvent en relation des groupes d'utilisateurs distincts et monétisent l'attention, les données ou les transactions. Dans le cadre des marchés bifaces, les plateformes ne sont pas des lieux passifs ; elles conçoivent des prix, des règles et des interfaces qui influencent la participation et le comportement (Rochet & Tirole, 2003, 2006). Les systèmes de recommandation modernes peuvent être considérés comme une couche supplémentaire de la conception des plateformes : ils mettent en œuvre une forme d'intermédiation algorithmique qui façonne l'attention et les choix au sein de la plateforme. La question de savoir si cette intermédiation améliore le bien-être dépend des incitations et des contraintes de la plateforme, ainsi que de la manière dont les utilisateurs traitent les informations qu'ils reçoivent.

Deux sources de friction sont au cœur de cette thèse. La première est le risque lié à la confidentialité en tant qu'attribut de l'ensemble des choix. Les avantages des services numériques (personnalisation et coûts de recherche réduits) s'accompagnent souvent d'une collecte de données et d'une exposition à une utilisation abusive ou à des fuites d'informations personnelles. La deuxième friction est l'opacité et l'asymétrie informationnelle des systèmes algorithmiques. Les utilisateurs observent rarement le fonctionnement d'un algorithme, l'objectif qu'il optimise et la manière dont leurs données sont intégrées dans le processus entre les entrées et les sorties (Eslami et al., 2015). Ces derniers doivent décider s'ils suivent les recommandations et s'ils délèguent les décisions en fonction de connaissances imparfaites.

En réponse à l'opacité et aux préoccupations de redevabilité, les régulateurs et les entreprises encouragent de plus en plus les interventions en faveur de la transparence : explications, divulgations et informations « claires et pertinentes » sur les systèmes automatisés. Dans l'Union européenne, ce vocabulaire est désormais explicite dans la loi sur l'IA, qui accorde aux personnes concernées le droit d'obtenir des explications claires et pertinentes sur le rôle d'un système d'IA à haut risque dans une procédure décisionnelle individuelle (article 86 ; considérant 171) (European Parliament & Council of the European Union, 2024). Cependant, il n'est pas évident que les explications puissent modifier de manière fiable les comportements, la confiance ou l'acceptation. Les explications peuvent mettre à jour les croyances sans pour autant influencer les choix ; elles peuvent fonctionner pour certains utilisateurs et pas pour d'autres ; et elles peuvent être dominées par des incitations, des chemins par défaut et des contraintes d'attention. Cette thèse traite donc de la transparence et de l'explicabilité non pas comme des objectifs normatifs, mais comme des mesures empiriques dont l'efficacité peut être mesurée.

À travers quatre essais, je combine des expériences en laboratoire, un essai contrôlé randomisé et une expérience en ligne à grande échelle afin d'identifier quand et comment les risques liés à la confidentialité et les interventions en matière de transparence par le biais d'explications claires et pertinentes influencent les comportements, les croyances et l'acceptation des algorithmes. La principale conclusion de la thèse est cohérente dans tous les contextes : la transparence et les explications peuvent influencer les croyances, mais elles ne constituent pas un levier universel pour modifier les comportements ou l'acceptation. Les

incitations, le contexte d'utilisation et l'apprentissage social sont souvent déterminants.

De l'*Homo Economicus* à l'*Homo Numericus*. L'économie modélise traditionnellement les agents comme des décideurs confrontés à des contraintes et à l'incertitude. Deux grandes avancées ont enrichi cette description. Premièrement, l'économie de l'information a formalisé la manière dont l'asymétrie de l'information et l'imperfection des signaux façonnent les marchés et les contrats (Stigler, 1961; Akerlof, 1970; Spence, 1973; Stiglitz, 1975; Grossman & Stiglitz, 1980). Ensuite, l'économie comportementale a documenté les écarts systématiques par rapport à la rationalité totale, tels que la rationalité limitée, l'heuristique et l'attention limitée (Simon, 1955; Tversky & Kahneman, 1974; Kahneman & Tversky, 1979; Gigerenzer & Goldstein, 1996; Gigerenzer & Gaissmaier, 2011). Le capitalisme numérique combine et amplifie ces deux thèmes.

Les systèmes algorithmiques interviennent directement dans le processus décisionnel. Au niveau du « menu », ils sélectionnent et classent les options, façonnant ainsi les ensembles de considérations, la variété perçue et les coûts de recherche. Au niveau des « croyances », ils génèrent des signaux et des récits sur ce qui est pertinent, sûr, populaire ou optimal. Un utilisateur qui choisit un plan d'épargne, par exemple, n'est pas confronté à l'ensemble des portefeuilles disponibles ; il est confronté à une recommandation, à un ensemble restreint de déviations et à une interface qui encadre le risque et le rendement. Un auditeur qui explore la musique n'est pas confronté à l'univers des morceaux ; il est confronté à une liste classée de suggestions et à un flux de « découvertes » qui concentre son attention. Dans les deux cas, l'algorithme n'est pas seulement un outil ; il modifie l'environnement de choix en restreignant et en ordonnant l'ensemble des choix et en façonnant les signaux sur lesquels les individus fondent leurs décisions.

Dans le même temps, les plateformes extraient des informations à partir des comportements : les clics, les recherches et la consommation deviennent des données. Sur de nombreux marchés, ces données améliorent la mise en correspondance et la personnalisation ; elles peuvent également permettre la persuasion, la discrimination par les prix ou l'exposition à des atteintes à la vie privée (Acquisti et al., 2016). Du point de vue de l'utilisateur, le problème de décision devient complexe. L'utilisateur choisit le contenu, les produits ou les portefeuilles, mais choisit aussi implicitement les informations à révéler et le degré de délégation à accepter. C'est l'une des raisons pour lesquelles le terme *Homo Numericus* est utile : il décrit un décideur dont les préférences sont filtrées par une médiation algorithmique et dont les actions sont également des signaux informationnels.

Ce terme recouvre plusieurs caractéristiques étroitement liées. Les choix sont influencés par des systèmes de recommandation, des classements et des outils prédictifs qui filtrent les options efficaces et les informations pertinentes. Parallèlement, les décisions dépendent des données : les actions individuelles génèrent des traces qui sont réinjectées dans le système et conditionnent les recommandations futures.

Cette interaction se caractérise par son opacité et son asymétrie informationnelle, car les utilisateurs ont rarement connaissance des objectifs, des contraintes ou de la logique interne du système. Enfin, les résultats sont sensibles au contexte, car l'attention et la délibération dépendent de la conception de l'interface et des conditions dans lesquelles l'information est rencontrée.

Cette thèse ne prétend pas que la modélisation économique standard soit devenue obsolète. Elle soutient plutôt que la médiation algorithmique est une technologie de l'information au sens précis de l'économie de l'information : elle modifie les signaux, les croyances et la correspondance entre les préférences et les choix. Empiriquement, cela implique que les conséquences des systèmes algorithmiques sur le bien-être ne peuvent être déduites uniquement à partir des performances techniques. Elles dépendent de la manière dont les utilisateurs perçoivent le risque, traitent les explications, actualisent leurs connaissances à partir de leur expérience et décident de déléguer ou non.

Les algorithmes en tant qu'intermédiaires de l'information : signaux, architecture des choix et conception de l'information

Une manière utile d'interpréter la médiation algorithmique consiste à passer par le traitement de l'information. Dans les modèles classiques, les agents sont confrontés à l'incertitude et choisissent leurs actions en fonction de signaux. Cependant, lorsque l'attention est limitée, le *format*, la *saillance* et le *coût cognitif* de l'information deviennent des primitives économiques (Sims, 2003; Chetty et al., 2009; Matějka & McKay, 2015; Gabaix, 2019). Les interfaces numériques façonnent directement ces primitives : elles déterminent ce qui est saillant, ce qui se cache derrière les clics et le coût d'acquisition d'informations supplémentaires.

Les algorithmes en tant que signaux et architecture de choix. Tout d'abord, les résultats algorithmiques sont des signaux : les recommandations transmettent des informations (ou du moins semblent transmettre des informations) sur la qualité de la correspondance, le risque, la popularité ou l'approbation sociale. Les utilisateurs peuvent considérer ces résultats comme des conseils d'experts, des statistiques récapitulatives ou des indices persuasifs. La littérature sur l'aversion et l'appréciation des algorithmes souligne que les réponses aux conseils algorithmiques dépendent non seulement de leur exactitude, mais aussi de la compétence perçue, de l'équité et de la saillance des erreurs (Dietvorst et al., 2015; Logg et al., 2019). En d'autres termes, une même recommandation peut être interprétée différemment en fonction des croyances sur le système et de l'expérience antérieure.

Deuxièmement, les interfaces algorithmiques constituent une architecture de choix : elles organisent les menus et encadrent les options, améliorant parfois les résultats, parfois influençant les comporte-

ments (Thaler & Sunstein, 2008). Dans un robot-conseiller, la recommandation n'est pas simplement une information ; c'est le chemin par défaut à suivre lors de l'intégration. Dans une plateforme de streaming, la liste de recommandations n'est pas seulement une prévision des goûts, c'est la principale porte d'entrée vers la consommation. Ainsi, le rôle économique des algorithmes est à la fois informationnel et institutionnel : ils sont à la fois des signaux et des mécanismes qui répartissent l'attention.

Ce double rôle met en évidence l'utilité de distinguer la « compréhension » du « comportement ». Les utilisateurs peuvent comprendre ce qu'un algorithme « utilise », mais continuer à le suivre parce que l'interface rend tout écart coûteux ; à l'inverse, les utilisateurs peuvent s'écarter même s'ils comprennent le système, s'ils perçoivent un désalignement ou si les enjeux sont importants. Les conceptions empiriques de cette thèse sont construites pour mesurer les deux.

Les algorithmes comme moyen de persuasion et de conception de l'information. Un autre aspect est que les systèmes algorithmiques peuvent être conçus non seulement pour prédire, mais aussi pour influencer. Les plateformes peuvent être incitées à orienter les comportements vers un engagement plus important, des marges plus élevées ou une exposition moindre au risque. D'un point de vue économique, cela peut être lié à la conception de l'information et à la persuasion : l'émetteur (plateforme/algorithme) choisit les signaux à fournir au récepteur (utilisateur) afin de façonner ses actions (DellaVigna & Gentzkow, 2010; Kamenica & Gentzkow, 2011; Bergemann & Morris, 2019). Il est important de noter que la persuasion ne nécessite pas de tromperie ; elle peut résulter d'une divulgation sélective, d'un cadrage et d'un classement.

Cette perspective est importante pour la transparence. Si une explication fait elle-même partie de la politique de communication de l'expéditeur, alors les explications peuvent être stratégiques : elles peuvent révéler certains aspects de l'algorithme tout en occultant d'autres. Les utilisateurs peuvent également anticiper une telle stratégie et ne pas tenir compte des explications en conséquence. Cette tension motive le cadrage principal-agent développé ci-dessous : les explications sont des tentatives de réduire l'asymétrie, mais elles sont fournies par un agent qui peut ne pas être parfaitement aligné.

La confidentialité comme attribut risqué des options. La confidentialité pose un problème décisionnel particulier, car elle implique souvent des préjudices probabilistes dont l'ampleur est incertaine et les conséquences différées. Les individus échangent des avantages immédiats (commodité, personnalisation) contre des coûts futurs incertains (fuites, utilisation abusive, discrimination) (Acquisti et al., 2016; Acquisti & Grossklags, 2005). En ce sens, la confidentialité peut être modélisée comme un attribut risqué associé à l'option elle-même.

Une complication supplémentaire réside dans le fait que le risque lié à la confidentialité est intrinsèquement interpersonnel dans de nombreux contextes numériques. Les conséquences dépendent de qui

observe les informations et de la manière dont cette personne réagit, et les individus peuvent prendre conscience des risques non seulement à travers leur propre expérience, mais aussi en observant les violations qui affectent les autres. Cet apprentissage par observation fait écho aux mécanismes mis en avant dans l'économie de l'information, notamment le comportement grégaire et les cascades informationnelles (Banerjee, 1992; Bikhchandani et al., 1992). Dans les contextes numériques où les événements graves sont rares pour un utilisateur individuel mais marquants au niveau de la population, l'apprentissage par procuration peut vraisemblablement dominer l'apprentissage direct. L'une des contributions de la thèse est de quantifier ces canaux dans un environnement contrôlé.

Interventions en matière de transparence : ce que les explications peuvent (et ne peuvent pas) faire. En réponse aux préoccupations liées à l'opacité et à la responsabilité, les régulateurs et les entreprises encouragent de plus en plus la transparence par le biais d'explications. La littérature sur l'intelligence artificielle explicable (XAI) a développé diverses méthodes pour interpréter les prédictions des modèles (Doshi-Velez & Kim, 2017; Adadi & Berrada, 2018; Arrieta et al., 2020), et des techniques largement utilisées telles que LIME et les décompositions basées sur Shapley sont devenues des références (Ribeiro et al., 2016b; Lundberg & Lee, 2017). Cependant, l'impact économique des explications dépend de leur capacité à modifier de manière significative les croyances, les comportements et la délégation.

Les explications peuvent (i) améliorer la compréhension perçue, (ii) actualiser les croyances concernant les critères et les objectifs, (iii) influencer la perception de l'équité ou de la responsabilité, puis (iv) modifier l'adhésion ou la volonté de déléguer. À l'inverse, elles peuvent être ignorées en raison d'une attention limitée, mal comprises en raison de leur complexité, ou même réduire la confiance si elles mettent en évidence des compromis controversés. Il est important de noter que les effets de la transparence peuvent être hétérogènes, car le coût du traitement des explications diffère selon les utilisateurs et les contextes. C'est l'une des raisons pour lesquelles la mesure de l'engagement vis-à-vis des explications est essentielle dans cette thèse.

Une vision principal-agent de la médiation algorithmique

L'une des hypothèses centrales de cette thèse est que de nombreuses interactions avec les systèmes algorithmiques s'inscrivent naturellement dans le cadre d'un problème principal-agent. Un mandant délègue une tâche à un mandataire dont les actions sont imparfaitement observables et dont les objectifs peuvent ne pas correspondre entièrement à ceux du mandant (Ross, 1973; Jensen & Meckling, 1976; Holmström, 1979; Grossman & Hart, 1983). La délégation crée une asymétrie informationnelle et ouvre la voie à un aléa moral ou à un comportement stratégique lorsque la surveillance est limitée.

Les systèmes algorithmiques, et les organisations qui les déploient, s'inscrivent directement dans cette

structure. Les individus délèguent la recherche, le filtrage et l'évaluation au système et, dans certains cas, des recommandations à haut risque (par exemple, les profils de risque des portefeuilles). Cette délégation se fait dans l'opacité, car les utilisateurs n'observent généralement que les résultats (c'est-à-dire les scores ou les recommandations) et non la fonction objective, les contraintes ou les données d'entraînement qui les produisent. En conséquence, la délégation peut s'accompagner d'un désalignement : le système peut être optimisé pour l'engagement, la monétisation, la fidélisation, la conformité ou la gestion des risques plutôt que pour le bien-être de l'utilisateur.

Cette perspective devient particulièrement utile lorsque l'on distingue les asymétries d'information *ex ante* et *ex post*. Une distinction qui structure le rôle des explications tout au long de la thèse.

Asymétrie *ex ante* : décider de déléguer ou de divulguer. Avant d'utiliser un système algorithmique, les individus manquent généralement d'informations sur la manière dont le système utilise leurs données, ce qu'il optimise et la distribution des résultats auxquels ils peuvent s'attendre. En termes de relation mandant-mandataire, le mandant choisit de déléguer sans observer le type, les incitations ou la technologie du mandataire. Cela s'apparente à une sélection adverse : les utilisateurs ne peuvent pas distinguer complètement les systèmes compétents et alignés des systèmes extractifs ou mal gérés.

Sur de nombreux marchés, la divulgation peut partiellement résoudre ces asymétries (Grossman, 1981; Milgrom, 1981). Cependant, il est peu probable que ces asymétries soient complètement éliminées dans le contexte algorithmique : les divulgations peuvent être complexes, l'attention peut être limitée et les entreprises peuvent être incitées à fournir des informations minimales ou stratégiques plutôt que des informations claires améliorant le bien-être (Gabaix & Laibson, 2006; McDonald & Cranor, 2008; Ben-Shahar & Schneider, 2014). En conséquence, la délégation et la divulgation *ex ante* (partage de données, acceptation des conditions, intégration) se produisent souvent dans le cadre de croyances imparfaites façonnées par la réputation, les indices d'interface et l'expérience antérieure.

Un point connexe est que l'opacité *ex ante* peut interagir avec la confidentialité. Si les pratiques de collecte de données sont difficiles à observer, les utilisateurs ne peuvent pas évaluer avec précision le coût de la délégation en termes de confidentialité. Cela soulève un problème classique d'information : lorsque la dimension qualitative (protection de la vie privée) est difficile à observer, la concurrence peut ne pas offrir le niveau de confidentialité souhaité par la société, et des exigences de divulgation peuvent être proposées comme remèdes. Cette thèse fournit des preuves pertinentes à cette logique en étudiant la manière dont les individus réagissent lorsque le risque pour la vie privée est explicité et lorsque le fonctionnement des algorithmes est expliqué.

Asymétrie *ex post* : décider de suivre ou non les recommandations. Même après qu'un individu ait adopté un système, un deuxième problème d'information se pose : le mandant doit-il suivre

la recommandation de l'agent ? En finance, cela se traduit par l'adhésion à un profil de risque recommandé ; en musique, cela se traduit par le choix d'accepter les recommandations ou d'explorer de manière indépendante ; dans tous les domaines, cela se traduit par la volonté de s'en remettre à des systèmes automatisés.

L'asymétrie *ex post* est importante car le mandant ne peut pas facilement vérifier le fondement de la recommandation ni évaluer les alternatives contrefactuelles. La communication entre le mandant et l'agent est donc essentielle. Dans les modèles de communication stratégique, la transmission d'informations est limitée par les incitations et la crédibilité (Crawford & Sobel, 1982). Si les objectifs de l'expéditeur divergent de ceux du destinataire, les messages peuvent, au mieux, n'être que partiellement informatifs. Dans les contextes algorithmiques, les explications peuvent être interprétées comme des messages sur le fondement des recommandations. Si les utilisateurs perçoivent un désalignement, ils peuvent ignorer les explications ; s'ils perçoivent un alignement, les explications peuvent renforcer la confiance.

C'est une façon de comprendre pourquoi les explications peuvent fonctionner dans certains contextes et échouer dans d'autres : leur effet dépend de la confiance de base, des incitations perçues et des coûts de vérification. Ces conditions varient vraisemblablement selon les domaines (finance ou divertissement), les utilisateurs (expérimentés ou novices) et les contextes d'utilisation (ordinateur de bureau ou mobile).

Quel est le rôle d'une explication en termes économiques ? Cette thèse traite les explications comme des interventions pouvant influencer les résultats par le biais de trois canaux. Premièrement, les explications peuvent agir par le biais d'un canal d'information en améliorant la compréhension des utilisateurs quant à l'utilisation du système et aux raisons pour lesquelles il produit un résultat donné. Deuxièmement, elles peuvent agir par le biais d'un canal d'agence en influençant l'alignement et la confiance perçus : les explications peuvent rassurer les utilisateurs sur le fait que l'objectif de l'agent est compatible avec le leur, ce qui renforce la délégation et l'adhésion. Troisièmement, les explications interagissent avec les frictions décisionnelles par le biais de l'attention et des coûts cognitifs. Le traitement d'une explication étant coûteux, son effet dépend de l'attention que lui accordent les utilisateurs (Sims, 2003; Matějka & McKay, 2015).

Une contribution empirique clé de cette thèse consiste à identifier où cette chaîne réussit et où elle se rompt. Les croyances peuvent évoluer tandis que le comportement reste inchangé ; les effets comportementaux peuvent se concentrer parmi les utilisateurs qui s'engagent réellement dans l'explication ; et le contexte (appareil, expérience antérieure, enjeux) peut dominer. En termes de principal-agent, les explications peuvent améliorer la surveillance ou la transparence perçue, mais cela ne modifie pas automatiquement le contrat effectif si les incitations, les défauts ou les coûts de changement maintiennent le comportement ancré.

Positionnement dans la littérature : information, délégation, confidentialité et transparence. Cette thèse contribue à une littérature interdisciplinaire en pleine expansion, mais ses questions fondamentales sont essentiellement économiques : comment les structures d'information créées par les plateformes influencent-elles la prise de décision, et les interventions en matière de transparence peuvent-elles atténuer l'asymétrie informationnelle lorsque les individus délèguent à des systèmes algorithmiques dans des environnements numériques ?

Tout d'abord, elle s'appuie sur l'économie de l'information, où les coûts de recherche et l'acquisition d'informations déterminent les résultats du marché (Stigler, 1961) et où la distribution et la divulgation d'informations sont essentielles pour les incitations et l'équilibre (Akerlof, 1970; Spence, 1973; Stiglitz, 1975; Grossman & Stiglitz, 1980). Les plateformes numériques réduisent certains coûts de recherche tout en introduisant de nouvelles frictions informationnelles : les utilisateurs voient des signaux sélectionnés, mais n'observent pas comment ils sont produits.

Deuxièmement, elle est liée à la délégation et à la théorie de l'agence. La théorie classique de l'agence formalise la manière dont les actions et les informations cachées façonnent les contrats et la conception organisationnelle (Ross, 1973; Holmström, 1979; Grossman & Hart, 1983). Dans les contextes algorithmiques, l'« agent » est un système intégré à une organisation et à une plateforme, et la surveillance prend la forme de transparence et d'interprétabilité. Les modèles de communication stratégique soulignent que la transmission d'informations est limitée par les incitations et la crédibilité (Crawford & Sobel, 1982), ce qui fournit un langage naturel pour interpréter les explications comme des messages sur le raisonnement algorithmique.

Troisièmement, elle contribue à l'économie de la vie privée, qui met l'accent sur les compromis entre les avantages de l'utilisation des données et les dommages potentiels, et souligne pourquoi les choix en matière de vie privée peuvent diverger des préoccupations exprimées (Acquisti & Grossklags, 2005; Acquisti et al., 2016). Le premier essai de cette thèse fournit un environnement de référence où le risque pour la vie privée est explicite et immédiat, aidant à interpréter le comportement dans des situations réelles où les probabilités sont ambiguës et les conséquences différées.

Enfin, elle fait référence à la littérature émergente sur l'explicabilité et l'acceptation des systèmes algorithmiques, où les explications sont censées améliorer la confiance et la légitimité, mais où les preuves sont mitigées (Cramer et al., 2008; Doshi-Velez & Kim, 2017; Arrieta et al., 2020; Aysolmaz et al., 2023). En mesurant les changements de croyances, les résultats comportementaux, l'engagement et l'hétérogénéité, cette thèse clarifie quand les explications sont importantes et quand elles ne le sont pas.

Deux objets centraux : le risque pour la vie privée et les « explications claires et pertinentes »

Cette thèse se concentre sur deux objets qui sont au cœur des choix médiés par le numérique et de la réglementation contemporaine : le **risque pour la vie privée** et les **explications claires et pertinentes**. L'expression « explications claires et pertinentes » est utilisée ici dans l'esprit des récentes exigences de transparence de l'Union européenne pour les systèmes algorithmiques, et cette thèse l'étudie de manière empirique en se demandant si les explications influencent les croyances, les comportements et la volonté de déléguer.

Risque pour la vie privée : évaluation, importance et apprentissage social. La littérature sur la vie privée montre que les individus expriment souvent de vives inquiétudes lorsqu'ils divulguent des informations importantes dans la pratique, un phénomène parfois décrit comme un « paradoxe de la vie privée » (Acquisti et al., 2016). Une interprétation est que les préférences en matière de confidentialité sont instables et dépendent du contexte ; une autre est que les individus sont confrontés à des croyances bruitées et à une attention limitée, tandis que les risques réels sont ambigus et les conséquences différées. La confidentialité peut donc être sous-pondérée par rapport aux gains immédiats, même lorsque les individus y accordent de l'importance.

Cette thèse apporte une contribution en étudiant le risque lié à la vie privée dans un contexte où les probabilités sont explicites et les résultats immédiats : une tâche d'investissement dynamique où une option offre un rendement attendu plus élevé, mais comporte une probabilité connue de fuite de données dans laquelle certaines informations financières sont révélées à des pairs. Cette conception isole un compromis de référence entre l'argent et la vie privée dans un contexte de risque transparent. Un contexte qui aide à interpréter pourquoi la vie privée peut être sous-évaluée dans le domaine, où les probabilités sont rarement connues.

En outre, les convictions en matière de confidentialité peuvent être façonnées par l'expérience indirecte. Les utilisateurs observent les violations qui touchent d'autres personnes, lisent des articles sur les fuites ou déduisent les risques à partir des résultats de leurs pairs. La thèse accorde donc une attention particulière au rôle de l'apprentissage social dans la prise de décision liée à la confidentialité (Banerjee, 1992; Bikhchandani et al., 1992). Cela est non seulement pertinent d'un point de vue descriptif, mais aussi d'un point de vue politique : si l'apprentissage social est un canal majeur, alors les interventions en matière de confidentialité qui s'appuient sur l'expérience individuelle (ou sur des divulgations discrètes et individualisées) peuvent avoir des effets plus faibles que les interventions qui mettent en évidence les risques sur le plan social.

« **Explications claires et pertinentes** » : un idéal politique aux effets comportementaux incertains. Les exigences en matière de transparence se sont multipliées dans le domaine de la réglementation numérique et de la gouvernance des plateformes, souvent dans le but de fournir des explications « claires et pertinentes » : des comptes rendus succincts et accessibles sur le fonctionnement des systèmes et les facteurs importants. D'un point de vue économique, les explications visent à réduire l'asymétrie informationnelle dans les interactions entre mandants et mandataires, tant *ex ante* (compréhension de la logique du système et de l'utilisation des données) qu'*ex post* (compréhension du fondement d'une recommandation particulière).

Cependant, il n'est pas évident que les explications se traduisent par un changement de comportement. Les explications peuvent améliorer la compréhension et la confiance perçues sans modifier les choix, en particulier lorsque les incitations, les paramètres par défaut et les contraintes de temps prédominent. Les données sur la transparence et la confiance dans les systèmes de recommandation sont mitigées, et les effets peuvent dépendre des objectifs des utilisateurs et de la répartition de leur attention (Cramer et al., 2008; Eslami et al., 2015; Kizilcec, 2016; Miller, 2019; Poursabzi-Sangdeh et al., 2021). De manière plus générale, la divulgation peut échouer lorsque les informations sont complexes ou coûteuses à traiter, même lorsqu'elles sont techniquement « disponibles » (Sims, 2003; McDonald & Cranor, 2008; Ben-Shahar & Schneider, 2014; Matějka & McKay, 2015; Gabaix, 2019). Cette thèse évalue donc les explications en tant qu'interventions et mesure leur impact sur (i) les croyances, (ii) le comportement et (iii) l'acceptation et la délégation.

Stratégie empirique et plan de la dissertation

Pourquoi des expériences ? Les expériences sont particulièrement adaptées à cet objectif pour trois raisons. Premièrement, elles permettent de manipuler directement l'environnement informationnel : probabilités de fuite, contenu explicatif, alignement des recommandations ou format de divulgation. Deuxièmement, elles aident à séparer la mise à jour des croyances du comportement motivé par des incitations : on peut observer si les explications modifient la compréhension même lorsque les choix ne changent pas. Troisièmement, les expériences révèlent l'hétérogénéité et les mécanismes, tels que l'engagement différentiel avec les explications ou l'apprentissage social à partir des résultats des autres.

Cette thèse combine des conceptions complémentaires :

- Un cadre de laboratoire avec des choix répétés présentant un risque pour la vie privée et une divulgation sociale immédiate (chapitre 1),
- Un essai contrôlé randomisé sur le terrain intégré dans une interface financière réelle (chapitre 2),

- Une tâche d'écoute contrôlée en laboratoire étudiant la découverte et l'engagement avec les explications (chapitre 3),
- Une expérience en ligne sur un échantillon représentatif mesurant l'acceptabilité des systèmes algorithmiques (chapitre 4).

Un thème méthodologique est la mesure conjointe des **choix** et des **croyances**. Les choix sont importants pour le bien-être et les politiques, mais les croyances sont cruciales pour les mécanismes. Si les explications modifient les croyances mais pas les comportements, cela suggère l'existence de contraintes ou de frictions (valeurs par défaut, attention, incitations). Si les comportements changent sans modification claire des croyances, cela suggère l'existence d'autres canaux (signaux de confiance, responsabilité perçue ou simple saillance). Les designs empiriques de cette thèse sont construits pour distinguer ces cas.

Un fil conducteur conceptuel unificateur. Bien que chaque essai soit indépendant, cette thèse s'articule autour d'une logique commune : les informations sur les risques liés à la vie privée ou sur les systèmes algorithmiques peuvent modifier les croyances et les attitudes, et ces changements peuvent se traduire par des choix observables et une volonté de déléguer. Cette représentation est délibérément stylisée. Les individus partent de croyances antérieures, traitent les informations de manière imparfaite en raison de contraintes d'attention, et leurs actions peuvent générer des expériences qui alimentent leurs croyances et attitudes ultérieures. L'objectif empirique est donc d'identifier où la transmission de l'information aux croyances, et des croyances au comportement, est forte ou faible, et comment elle varie selon les contextes. Dans de nombreux environnements numériques, les croyances peuvent changer tandis que le comportement reste ancré par des incitations, des paramètres par défaut ou une attention limitée. Cette thèse mesure donc à la fois les résultats en matière de croyances et les résultats en matière de choix, et documente l'hétérogénéité entre les utilisateurs et les environnements.

Plan de la thèse : les quatre essais.

Chapitre 1 - *Trading Returns for Privacy: Experimental Evidence from Financial Data Leaks.*

Question. Comment les individus échangent-ils des rendements financiers attendus contre des probabilités explicites de fuites de données personnelles, et comment actualisent-ils leurs connaissances à partir de leurs propres expériences en matière de fuites et de celles des autres ?

Conception. Dans le cadre d'une tâche d'investissement dynamique en laboratoire, une option offre un rendement attendu plus élevé, mais comporte une probabilité connue que certaines informations financières soient révélées à des pairs ; l'autre option garantit la confidentialité. La tâche est répétée plusieurs fois, ce qui permet d'estimer la sensibilité aux écarts de rendement et aux probabilités de

fuite tout en suivant l'apprentissage dynamique à partir des résultats obtenus. La nature sociale de la divulgation permet de distinguer clairement entre (i) l'apprentissage tiré de sa propre expérience de fuite et (ii) l'apprentissage tiré de l'observation des fuites subies par d'autres.

Principales conclusions. Les choix d'allocation répondent beaucoup plus aux écarts de rendement qu'aux probabilités de fuite. Les expériences personnelles de fuite ont des effets modestes et de courte durée, tandis que l'observation d'un plus grand nombre de fuites parmi les pairs réduit considérablement les choix ultérieurs de l'option présentant un risque pour la confidentialité. Dans cet environnement, l'apprentissage social des résultats en matière de confidentialité est donc un facteur clé du comportement.

Interprétation. Même dans le cas d'un risque pour la vie privée inhabituellement transparent et immédiat, les incitations financières prédominent. Ce point de référence aide à expliquer pourquoi les choix protégeant la vie privée peuvent être faibles sur les marchés naturels où les risques sont opaques et différés (Acquisti et al., 2016). De plus, le rôle important des résultats des pairs suggère que les croyances et les comportements en matière de vie privée peuvent être façonnés de manière disproportionnée par des événements marquants et des informations sociales plutôt que par l'expérience personnelle.

Chapitre 2 - *Algorithmic Transparency and Portfolio Choices: Field Evidence.*

Question. La transparence algorithmique améliore-t-elle l'adhésion aux décisions financières à haut risque ?

Cadrage principal-agent. Le robo-conseil est un problème de délégation canonique : l'utilisateur (le principal) reçoit une recommandation de profil de risque d'un système algorithmique (l'agent) dont le mappage entre les entrées et les sorties est opaque. Les explications visent à réduire l'asymétrie informationnelle et à améliorer l'alignement perçu, augmentant ainsi l'adhésion et réduisant les écarts coûteux.

Conception. Dans le cadre d'un essai contrôlé randomisé intégré à l'interface d'intégration d'un robot-conseiller français de premier plan, $N = 3,856$ utilisateurs sont assignés de manière aléatoire pour recevoir (ou ne pas recevoir) des explications graphiques, basées sur leur profil, des facteurs déterminants de leur score de risque recommandé. L'essai porte sur 4 645 contrats d'épargne signés. La conception se concentre sur les résultats qui reflètent la force de la délégation : acceptation du profil recommandé, écarts par rapport à cette recommandation et engagement lors de l'intégration.

Principales conclusions. Fournir des explications n'augmente pas l'acceptation du profil recommandé, ne réduit pas les écarts par rapport au profil de risque recommandé et n'augmente pas l'engagement en moyenne. Au contraire, le comportement varie systématiquement en fonction du contexte numérique et de l'expérience antérieure des utilisateurs (par exemple, appareil et statut de retour ou de première utilisation).

Interprétation. Ces résultats sont cohérents avec un contexte dans lequel les explications peuvent

réduire l'opacité sans modifier le contrat effectif. Si l'attention est limitée, la valeur marginale du traitement d'une explication peut être faible par rapport à la simple poursuite du processus. Si les utilisateurs ont des *a priori* forts (sur le risque, sur la plateforme ou sur leurs propres préférences), les informations supplémentaires peuvent ne pas modifier leur comportement. Le contexte numérique est important car il détermine le coût du traitement de l'information et la friction de la délibération (Sims, 2003; Agnew & Szykman, 2005). D'un point de vue plus large, ces résultats incitent à la prudence et invitent à ne pas supposer que les mesures de transparence se traduiront automatiquement par une meilleure adhésion ou par des choix améliorant le bien-être.

Chapitre 3 - *Algo-Rhythm Unplugged: Effects of explaining algorithmic recommendations on music discovery.*

Question. Les explications modifient-elles le comportement des utilisateurs dans les systèmes de recommandation, et comment l'engagement avec les explications influence-t-il les effets ?

Conception. Dans le cadre d'une tâche d'écoute de 20 minutes, les participants reçoivent des recommandations qui correspondent à leurs préférences déclarées, qui ne correspondent pas à leurs préférences déclarées ou qui correspondent à leurs préférences déclarées avec une explication du fonctionnement du système. La conception mesure à la fois l'allocation d'écoute et les croyances concernant les critères du système de recommandation, et elle enregistre l'engagement envers l'explication à travers le temps passé à la consulter.

Principales conclusions. Les explications améliorent les croyances des participants concernant les principaux critères du système de recommandation, mais ne produisent pas de changement moyen significatif dans le temps d'écoute relatif. Une forte hétérogénéité apparaît : au sein du groupe traité, un engagement plus important avec les explications (temps passé à les consulter) est associé à un temps d'écoute plus long et à un changement dans le comportement de découverte.

Interprétation. Les explications peuvent fonctionner en actualisant les croyances, mais les effets comportementaux dépendent de l'attention endogène. Cela correspond à une attention limitée et à une inattention rationnelle : les utilisateurs ne tirent profit des explications que s'ils choisissent de les traiter (Sims, 2003; Matějka & McKay, 2015). Les résultats soulignent également pourquoi la « disponibilité » de la transparence n'est pas équivalente à l'« utilisation » de la transparence. En matière de politique et de conception, cela suggère que l'explicabilité est en partie un problème d'attention : la manière dont les explications sont intégrées dans le parcours de l'utilisateur peut être tout aussi importante que leur contenu.

Chapitre 4 - *Understanding, Trusting, Delegating: How Explanations Shape Public Support for ADM and AI Systems.*

Question. Les explications augmentent-elles de manière causale l’acceptabilité des systèmes algorithmiques dans tous les domaines ?

Conception. À partir d’un échantillon représentatif de trois grands pays d’Europe occidentale (France, Allemagne et Italie ; $N = 3,007$), les participants évaluent différents systèmes algorithmiques avec ou sans explication du fonctionnement de chaque système. Les résultats couvrent les croyances (connaissances, familiarité, exposition), les attitudes (confiance, équité, préoccupations en matière de confidentialité, utilité perçue) et la volonté de déléguer.

Principales conclusions. Les explications entraînent des gains dépendants du contexte en termes de croyances de base, d’attitudes et de volonté de déléguer, avec des effets plus faibles ou nuls dans certains domaines. Dans l’ensemble, les effets sont modestes et hétérogènes.

Interprétation. Les explications peuvent modifier l’acceptation, mais elles ne résolvent pas de manière uniforme les préoccupations relatives à la confidentialité, à l’équité et à la responsabilité. Ces préoccupations sont au cœur de la gouvernance des systèmes automatisés (Kroll et al., 2017; Aysolmaz et al., 2023). La dépendance au contexte suggère que les a priori des utilisateurs et les enjeux perçus de la délégation influencent la manière dont la transparence est interprétée.

Synthèse : ce que les quatre essais impliquent conjointement

Informations sur les risques ou les algorithmes → Croyances et attitudes → Choix et délégation.

Une conclusion empirique centrale est que la chaîne se rompt souvent entre les croyances et le comportement. Les explications peuvent être efficaces sur le plan cognitif ; elles peuvent améliorer la compréhension déclarée et parfois les attitudes sans modifier de manière fiable le comportement des personnes.

Les croyances peuvent être modifiées plus facilement que les choix. Dans l’expérience musicale, les explications améliorent de manière fiable la compréhension des critères de l’algorithme de recommandation, mais la répartition moyenne de l’écoute ne change pas de manière significative. Dans l’expérience interdomaines, les explications augmentent modestement la volonté de déléguer, principalement par le biais de changements d’attitude plutôt que par des mises à jour importantes des croyances factuelles. Dans le domaine des robots-conseillers, les explications ne modifient pas l’adhésion en moyenne. Ensemble, ces résultats suggèrent que la transparence est souvent plus efficace pour façonner les perceptions que pour modifier les choix, en particulier lorsque ceux-ci sont ancrés par des valeurs par défaut, des contraintes ou des a priori forts.

Ce schéma correspond à une interprétation simple basée sur l’attention. Le coût cognitif du traitement et de l’intégration d’une explication dans un choix peut dépasser sa valeur marginale, en particulier

lorsque la recommandation est déjà plausible ou lorsque la décision doit être prise rapidement. En termes d'inattention rationnelle, les explications sont des informations qui ont un coût ; leur importance dépend de la capacité de l'utilisateur à leur accorder une attention optimale (Sims, 2003; Chetty et al., 2009; Matějka & McKay, 2015; Gabaix, 2019). Les résultats hétérogènes de cette thèse (en particulier dans la tâche musicale) renforcent ce point de vue en montrant que les effets se concentrent parmi ceux qui y accordent leur attention.

Les incitations l'emportent sur les risques liés à la vie privée, même lorsque ces risques sont explicites. Le chapitre 1 fournit un point de référence : même lorsque le risque de fuite est explicite et se concrétise rapidement, les écarts de rendement ont un effet plus important que les probabilités de fuite. Cela ne signifie pas que la vie privée n'a pas d'importance, mais plutôt que les risques liés à la vie privée sont souvent secondaires par rapport aux incitations financières immédiates. Cela aide à comprendre pourquoi les mesures de protection de la vie privée qui reposent sur le choix individuel (opt-out, bannières de consentement) peuvent s'avérer peu efficaces lorsque les risques sont opaques et différés (Acquisti et al., 2016). Cela suggère également que mettre en avant la protection de la vie privée peut ne pas être suffisant si les incitations inhérentes à l'environnement de choix favorisent fortement les options présentant un risque pour la vie privée.

L'apprentissage social peut l'emporter sur l'expérience personnelle. Un résultat frappant du chapitre 1 est que le fait d'observer les fuites d'autrui réduit davantage les choix ultérieurs risqués pour la vie privée que le fait de subir soi-même une fuite. Cela fait écho aux modèles d'économie de l'information dans lesquels les agents déduisent le risque à partir des résultats d'autrui (Banerjee, 1992; Bikhchandani et al., 1992). Dans les environnements numériques, où les violations majeures sont rares pour un utilisateur individuel mais importantes au niveau de la population, les informations sociales peuvent être un facteur clé de la perception du risque et du comportement. Cela a des implications pratiques : les convictions en matière de confidentialité peuvent réagir fortement à des événements publics marquants, et les conséquences comportementales des incidents liés à la confidentialité peuvent se répercuter sur des utilisateurs qui n'ont pas été directement touchés.

Le contexte numérique est une variable économique. Le chapitre 2 souligne que le type d'appareil utilisé et l'expérience antérieure de la plateforme ne sont pas des détails de mise en œuvre ; ils influencent l'attention, la réflexion et le coût effectif du traitement des explications. Le contexte numérique fait donc partie intégrante de la structure de l'information. Les explications rencontrées sur un téléphone, lors d'une étape d'intégration précipitée ou par un utilisateur expérimenté peuvent avoir des effets très différents de ceux des mêmes explications rencontrées sur un ordinateur de bureau par un utilisateur inexpérimenté. Cela suggère une leçon plus large pour la politique numérique : la normalisation du contenu de la

transparence peut ne pas être suffisante si la transparence est rencontrée dans des conditions d'attention très différentes.

Les explications en tant qu'outils principal-agent : nécessaires mais insuffisantes. Dans tous les essais, les explications fonctionnent comme des remèdes partiels à l'asymétrie informationnelle dans les relations principal-agent. Pourtant, la théorie principal-agent prédit que la surveillance seule est rarement suffisante : les incitations et les contraintes ont leur importance (Holmström & Milgrom, 1991). Les résultats de cette thèse concordent avec ce point de vue. Les explications peuvent réduire l'opacité et faire évoluer les croyances, mais sans changements complémentaires en matière d'incitations, de gouvernance ou de responsabilité, elles ont souvent un impact comportemental limité. En bref, la transparence peut être utile, mais elle ne remplace pas l'alignement.

Contributions à l'économie. Cette thèse contribue à l'économie de la prise de décision et de l'information de quatre manières principales.

Une référence en matière de confidentialité dans un contexte de risque explicite et immédiat. En fournissant un cadre dans lequel le risque lié à la confidentialité est explicite et se réalise immédiatement, le chapitre 1 offre une référence claire pour l'évaluation et l'apprentissage de la confidentialité. Il identifie la prédominance des incitations financières et met en évidence l'apprentissage social comme un canal majeur qui façonne les choix liés à la confidentialité. Cela complète les travaux plus généraux sur l'économie de la confidentialité (Acquisti et al., 2016) en montrant ce qui se passe lorsque les frictions clés dans ce domaine, à savoir l'ambiguïté et le retard, sont délibérément minimisées.

Preuves expérimentales sur le terrain concernant les explications dans le domaine de la finance. Le chapitre 2 ajoute des preuves rares sur le terrain concernant les interventions en matière d'explicabilité dans les décisions importantes des ménages. L'effet moyen nul est instructif : il met en garde contre l'hypothèse selon laquelle les explications augmentent mécaniquement l'adhésion et suggère que le contexte et l'hétérogénéité sont essentiels pour l'évaluation. Les résultats encouragent également une vision plus centrée sur l'agence du conseil robotisé : les explications peuvent réduire l'asymétrie informationnelle sans modifier le comportement si d'autres caractéristiques de l'environnement délégué dominant.

Effets de transparence médiés par l'attention dans les plateformes de consommation. Le chapitre 3 montre que les explications peuvent modifier les croyances sans modifier le comportement en moyenne, tout en révélant une forte hétérogénéité liée à l'engagement. Cela confirme l'idée que les effets des explications dépendent de l'acquisition endogène d'informations et des contraintes d'attention (Sims, 2003; Matějka & McKay, 2015). Sur le plan méthodologique, cela incite à mesurer l'engagement et à interpréter la transparence comme une intervention en deux étapes : disponibilité et consommation.

Preuves causales intercontextuelles sur l’acceptabilité et la délégation. Le chapitre 4 fournit des preuves causales intercontextuelles sur la manière dont les explications façonnent les attitudes et la volonté de déléguer. La dépendance au domaine implique que des exigences de transparence uniformes peuvent avoir des effets inégaux et ne doivent pas être considérées comme fonctionnant de manière uniforme (Aysolmaz et al., 2023). Les résultats offrent également une approche de mesure structurée de l’acceptabilité, en séparant les croyances, les attitudes et la délégation.

Implications pour la politique et la gouvernance. Les exigences de transparence sont souvent justifiées par la protection des consommateurs et la mise en place d’une infrastructure de responsabilité. Cette thèse soutient leur pertinence institutionnelle tout en mettant en garde contre des promesses excessives en matière d’impact comportemental.

Premièrement, la transparence doit être considérée comme une intervention à évaluer empiriquement, et non comme une garantie de meilleures décisions. Dans de nombreux contextes, la divulgation peut échouer en raison de la complexité, de l’attention limitée ou de la conception stratégique (Grossman, 1981; Milgrom, 1981; Fung et al., 2007; Gabaix & Laibson, 2006; McDonald & Cranor, 2008; Dranove & Jin, 2010; Ben-Shahar & Schneider, 2014). Deuxièmement, l’adaptation est importante : la conception des explications doit refléter les préoccupations spécifiques au domaine et le contexte numérique dans lequel l’information est rencontrée. Troisièmement, la transparence ne remplace pas l’alignement. Même lorsque les explications augmentent l’acceptation, elles ne garantissent pas que les objectifs soient alignés ou que les préjudices soient évités ; une responsabilité et une gouvernance plus larges restent nécessaires (Kroll et al., 2017).

Une dernière implication concerne les normes d’évaluation. Les effets des explications étant hétérogènes, les effets moyens peuvent être trompeurs pour les politiques. Si seul un sous-ensemble d’utilisateurs s’intéresse aux explications (comme dans le chapitre 3), le problème de conception pertinent devient alors : comment intégrer la transparence de manière à ce qu’elle soit réellement traitée par ceux qui en ont besoin ? Si les effets dépendent de l’appareil et de l’expérience (comme dans le chapitre 2), les obligations d’explication doivent être évaluées dans l’interface spécifique où elles seront mises en œuvre, et pas seulement sur le papier. En bref, la transparence devrait être évaluée dans le cadre d’un *environnement informationnel*.

Conclusion et structure de la thèse. Dans l’ensemble, cette thèse offre un compte rendu empirique de la prise de décision pour l’*Homo Numericus* à l’ère du capitalisme numérique. Elle montre que les explications peuvent réduire les asymétries informationnelles dans la relation principal-agent entre les utilisateurs et les systèmes algorithmiques, mais que la transparence n’est pas une solution universelle. Les incitations, les contraintes d’attention, le contexte d’utilisation et l’apprentissage social influencent

souvent le comportement plus fortement que les explications. Pour l'économie, cela implique que la médiation algorithmique doit être traitée comme une structure d'information. En outre, les politiques de transparence et de confidentialité doivent être évaluées de manière empirique et adaptées aux domaines et aux contextes d'utilisation, plutôt que de supposer qu'elles fonctionnent de manière uniforme.

Mehdi LOUAFI

QUATRE ESSAIS SUR L'HOMO NUMERICUS : LA PRISE DE DÉCISION À L'ÈRE DU CAPITALISME NUMÉRIQUE

Résumé : Cette thèse étudie l'Homo Numericus, dont les décisions économiques sont de plus en plus médiées par des plateformes numériques, et analyse comment des systèmes algorithmiques (recommandation, classement, personnalisation) façonnent la prise de décision tout en collectant des données personnelles. À travers quatre essais expérimentaux, j'identifie quand et comment les risques de confidentialité et les interventions de transparence via des explications significatives influencent les comportements, croyances et l'acceptation des recommandations algorithmiques. Le premier essai mesure l'arbitrage rendement-fuite de données : dans une tâche d'investissement, les allocations réagissent davantage aux écarts de rendement qu'aux probabilités de fuite ; les fuites personnelles ont un effet faible, tandis que l'observation de fuites chez des pairs réduit durablement la prise de risque en matière de confidentialité. Le deuxième essai vise à déterminer si la transparence algorithmique améliore l'adhésion aux recommandations dans des décisions financières à forts enjeux. L'essai contrôlé randomisé intégré à l'interface d'un robot-conseiller (4645 contrats), montre que des explications graphiques du score de risque n'augmentent pas l'adhésion ni l'engagement en moyenne, mais que les effets varient avec le contexte numérique et l'expérience des utilisateurs. Le troisième essai montre que, dans la recommandation musicale, les explications améliorent les croyances sur le fonctionnement du système sans effet moyen robuste sur l'écoute, avec une forte hétérogénéité selon l'attention portée aux explications. Le dernier essai montre que les explications génèrent une augmentation faible mais détectable de la volonté de déléguer à des algorithmes. Ces effets s'expliquent principalement par une amélioration des attitudes envers les systèmes, tandis que les effets directs et les changements de croyances jouent un rôle secondaire. Au total, la transparence et les explications modifient surtout les croyances et les attitudes : incitations, contexte d'usage et apprentissage social dominant souvent les comportements, ce qui plaide pour des politiques d'explicabilité et de confidentialité évaluées empiriquement et adaptées aux usages.

Mots-clés : prise de décision ; algorithmes/IA ; délégation ; données personnelles ; transparence ; économie expérimentale

FOUR ESSAYS ON HOMO NUMERICUS: DECISION-MAKING IN THE AGE OF DIGITAL CAPITALISM

Abstract: This dissertation studies Homo Numericus—individuals whose economic choices are increasingly mediated by digital platforms—and examines how algorithmic systems (recommendation, ranking, personalization) shape decision-making while collecting personal data. Across four experimental essays, I identify when and how privacy risks and transparency interventions through meaningful explanations affect behavior, beliefs, and acceptance of algorithms. Essay 1 quantifies the return–privacy trade-off in a laboratory investment task: allocations respond far more to return spreads than to leak probabilities; own leak experiences have small effects, while observing leaks among peers significantly reduces later privacy-risky choices. Essay 2 tests whether algorithmic transparency improves adherence in high-stakes financial decisions. The randomized controlled trial embedded in the onboarding interface of a robo-advisor (4,645 saving-plan contracts) shows that graphical explanations of the recommended risk score do not increase acceptance, reduce deviations, or raise engagement on average, but responses vary with users' digital context and prior platform experience. Essay 3 shows that, in music recommendation, explanations improve beliefs about how the system works without a robust average change in listening behavior, with strong heterogeneity by attention to the explanation. Essay 4 finds that explanations generate small but detectable increases in willingness to delegate to algorithms. These effects arise primarily through improved attitudes toward the systems, while direct effects and belief shifts play a secondary role. Overall, transparency mainly shifts beliefs and attitudes: incentives, context of use, and social learning often dominate behavior, implying that explainability and privacy policies should be tested empirically and tailored to domains and user contexts.

Keywords: decision-making ; algorithms/AI ; delegation ; personal data ; transparency ; experimental economics