



HAL
open science

Misplaced trust in AI: the explanation paradox and the human-centric path. A characterisation of the cognitive challenges to appropriately trust algorithmic decisions and applications in the financial sector

Astrid Bertrand

► **To cite this version:**

Astrid Bertrand. Misplaced trust in AI: the explanation paradox and the human-centric path. A characterisation of the cognitive challenges to appropriately trust algorithmic decisions and applications in the financial sector. Artificial Intelligence [cs.AI]. Institut Polytechnique de Paris, 2024. English. NNT : 2024IPPAT012 . tel-04661844

HAL Id: tel-04661844

<https://theses.hal.science/tel-04661844>

Submitted on 25 Jul 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



INSTITUT
POLYTECHNIQUE
DE PARIS

NNT : 2024IPPAT012

Thèse de doctorat



Misplaced trust in AI: the explanation paradox and the human-centric path. A characterisation of the cognitive challenges to appropriately trust algorithmic decisions and applications in the financial sector.

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

École doctorale n°626 Ecole doctorale de l'Institut Polytechnique de Paris (ED IP Paris)
Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Palaiseau, le 21 mai 2024, par

ASTRID BERTRAND

Composition du Jury :

Alexandre de Streel Professeur, Université de Namur	Président/Examineur
Fosca Giannotti Professeure, Scuola Normale Superiore di Pisa	Rapporteuse
Tim Miller Professeur, University of Queensland	Rapporteur
Nadia Boukhelifa Chargée de recherche, INRAE, Université Paris Saclay	Examinatrice
Winston Maxwell Professeur, i3, CNRS, Télécom Paris	Directeur de thèse
James R. Eagan Maître de conférences, LTCI, Télécom Paris	Co-directeur de thèse
Olivier Fliche Directeur du pôle Fintech-Innovation, ACPR, Banque de France	Invité

Abstract

Deep learning, the technology behind ChatGPT relies on a complex and massive network of mathematical operations. Although we know the math for each individual operation, we do not understand why the network as a whole produces the results we see. For most of "artificial intelligence" (AI) models¹, it is unclear why they behave the way they do, making it difficult to determine when they fail and if they have biases. This problem has led to a significant growth of research on explainability in recent years, which focuses on understanding the behaviour of machine learning models. However, there has been comparatively little exploration of how current explainability methods align with the requirements of highly regulated environments such as finance, taking into account human factors. In such contexts, the warranted, *i.e.* well-calibrated trust of customers and regulators in AI systems can be critical for achieving regulatory compliance. This thesis explores the potential of explainability to enable warranted trust in AI and help ensure compliance of AI-enhanced systems in financial applications.

¹ The term "artificial intelligence" (AI) encompasses these deep learning techniques as well as less complex machine learning models.

The first part explores the cognitive barriers related to the construction of explainable AI interfaces that promote appropriate levels of trust, through two detailed scoping literature reviews. In the first analysis, we present a heuristic map of the different cognitive biases to be taken into account in the design of explainability through the review of 38 research articles. We also detail the context in which these different biases were found, in particular the method of explicitation used and the types of users and tasks in which they appear. This study reveals an 'explanation paradox', where explanations intended to inform users may ultimately increase their confidence in untrustworthy AI models, which is undesirable. The second detailed scoping literature review of this thesis studies a corpus of 48 articles and provides a taxonomy of the different ways of interacting with explainability solutions. We identify three categories of interaction according to their role in the cognitive process of explanation: 'selective', 'mutable' or 'dialogic'. We also analyse the effects of these types of interaction on users. We find that interactive explanations improve the perceived usefulness and performance of the human+AI team, but that they take longer. Finally, we describe some little-explored avenues, such as measuring curiosity or learning.

The second part deals with the needs and effects of explanations in financial contexts. We conduct a controlled study with 256 participants in the context of online life insurance distribution, where there are already legal requirements for explanations, to compare the effect of several types

of explanation representation on user trust. We show that feature-based explanations did not significantly improve customers' understanding of the recommendation or their ability to perceive its inappropriateness, a result that is the opposite of what the law hoped to achieve. In addition, explanations in the form of dialogue increased users' trust in the recommendations made by the robo-advisor, sometimes to the detriment of the users themselves. This real-life scenario illustrates how explainability can prove insufficient to remedy information asymmetry in complex areas such as finance. Another study analyses supervisors' requirements for explainability solutions in the fight against money laundering and the financing of terrorism (AML-CFT). Through scenario-based workshops with 13 supervisors and 6 banking industry professionals, we describe the audit practices and the supervisor's socio-technical context. Combining observations from the workshops with an analysis of compliance requirements, we identify AML-CFT obligations that conflict with AI opacity. We then articulate supervisors' needs for model justification. We discuss the role of explanations as reliable evidence on which to base justifications.

The conclusion discusses the potential of explanations to manipulate user trust. We then review promising human-centered development paths for developing explainable AI interfaces that enhance user autonomy. These include personalising explanations, presenting a range of options rather than a single recommendation/explanation, stimulating user scepticism, and fostering user engagement, curiosity and learning. The role of explainability in mitigating regulatory tensions caused by the use of opaque AI models in AML-CFT is also examined.

Résumé

L'apprentissage profond, la technologie derrière ChatGPT, repose sur un réseau complexe et massif d'opérations mathématiques. Bien que nous connaissions les mathématiques de chacune de ces opérations, nous ne comprenons pas pourquoi le réseau dans son ensemble produit les résultats que nous voyons. Pour la plupart des modèles d'« intelligence artificielle » (IA), on ne sait pas pourquoi ils se comportent comme ils le font, ce qui rend difficile de déterminer quand ils peuvent se tromper et s'ils ont des biais. Ce problème a conduit à une croissance significative de la recherche sur l'explicabilité au cours des dernières années, qui se concentre sur la compréhension du comportement des modèles d'apprentissage automatique. Toutefois, la façon dont les méthodes actuelles d'explicabilité s'alignent sur les exigences d'environnements hautement réglementés tels que la finance, en tenant compte des facteurs humains, a été relativement peu explorée. Dans de tels contextes, la confiance justifiée, *i.e.* bien calibrée des clients et des régulateurs dans les systèmes d'IA peut être critique pour atteindre la conformité réglementaire. Cette thèse explore le potentiel de l'explicabilité pour permettre une confiance justifiée dans l'IA et pour aider à assurer la conformité des systèmes améliorés par l'IA dans les applications financières.

La première partie explore les obstacles cognitifs liés à la construction d'interfaces d'IA explicables et favorisant des niveaux de confiance appropriés, grâce à deux examens détaillés de la littérature. Dans une première analyse, nous présentons une carte heuristique des différents biais cognitifs à prendre en compte dans la conception de l'explicabilité grâce à l'examen de 38 articles de recherche. Nous détaillons aussi le contexte dans lequel ces différents biais identifiés ont été trouvés, notamment la méthode d'explicitation utilisée et les types d'utilisateurs et de tâches dans lesquels ils apparaissent. Cette étude révèle un « paradoxe de l'explication », où les explications destinées à informer les utilisateurs peuvent finalement accroître leur confiance dans des modèles d'IA non dignes de confiance, ce qui n'est pas souhaitable. La deuxième revue de littérature de cette thèse étudie un corpus de 48 articles et fournit une taxonomie des différentes façons d'interagir avec les solutions d'explicabilité. Nous déterminons trois catégories d'interaction en fonction de leur rôle dans le processus cognitif d'explication : « sélectif », « mutable » ou « dialogique ». Nous analysons également les effets de ces types d'interaction sur les utilisateurs. Nous constatons que les explications interactives améliorent l'utilité perçue et la performance de l'équipe humaine+AI, mais qu'elles prennent plus de temps. Enfin, nous

décrivons des pistes peu explorées, notamment la mesure de la curiosité ou de l'apprentissage.

La deuxième partie traite des besoins et des effets des explications dans les contextes financiers. Nous menons une étude contrôlée avec 256 participants dans le contexte de la distribution en ligne d'assurances-vie, où il existe déjà des exigences légales en matière d'explications, pour comparer l'effet sur la confiance des utilisateurs de plusieurs types de représentation d'explications. Nous montrons que les explications basées sur les caractéristiques n'amélioreraient pas de manière significative la compréhension de la recommandation par les clients ou leur capacité à percevoir son caractère inapproprié, un résultat qui est à l'opposé de ce que la loi espérait obtenir. En outre, les explications sous forme de dialogue augmentent la confiance des utilisateurs dans les recommandations du robot-conseiller, parfois au détriment des utilisateurs. Ce scénario réel illustre comment l'explicabilité peut se révéler insuffisante pour remédier à l'asymétrie de l'information dans des domaines complexes tels que la finance.

Une autre étude analyse les exigences des autorités de contrôle en matière de solutions d'explicabilité dans le cadre de la lutte contre le blanchiment d'argent et le financement du terrorisme (LCB-FT). Grâce à des ateliers basés sur des scénarios avec 13 superviseurs et 6 professionnels du secteur bancaire, nous décrivons les pratiques d'audit et le contexte sociotechnique du superviseur. En combinant les observations des ateliers avec une analyse des exigences de conformité, nous identifions les obligations en matière de LCB-FT qui entrent en conflit avec l'opacité de l'IA. Nous formulons ensuite les besoins des superviseurs en matière de justification des modèles. Nous discutons du rôle des explications en tant que preuves fiables sur lesquelles fonder les justifications.

La conclusion aborde le potentiel des explications pour manipuler la confiance des utilisateurs. Nous passons ensuite en revue les pistes de développement centrées sur l'humain qui sont prometteuses pour développer des interfaces d'IA explicables qui améliorent l'autonomie des utilisateurs. Ces pistes sont la personnalisation des explications, la présentation d'un éventail d'options plutôt que d'une seule recommandation/explication, la stimulation du scepticisme des utilisateurs, et la favorisation de l'engagement, de la curiosité et de l'apprentissage des utilisateurs. Le rôle de l'explicabilité dans l'atténuation des tensions réglementaires causées par l'utilisation de modèles d'IA opaques dans la LCB-FT est également examiné.

Acknowledgments

Work setup

I have been fortunate to be welcomed into a variety of working environments during my PhD. This thesis is a reflection of these many experiences.

My research was carried out as part of a PhD contract with Télécom Paris, an engineering school affiliated to the Institut Polytechnique de Paris. Since my arrival at Télécom, I have been part of the newly formed *Operational AI Ethics* (OpAIE) team, founded by Winston Maxwell, which explores interdisciplinary issues related to the societal impact of AI. As one of the first PhD students of this group, it has been fantastic to see it grow over the last three years—we are now over 6 PhD students and 3 full-time professors. Starting from the second year of my PhD, thanks to James Eagan, I joined the *Design, Interaction, Visualizations and Applications* (DIVA) group, the Human-Computer Interaction (HCI) team of Télécom Paris. Finding colleagues with whom to discuss HCI methods, conferences and best practices was invaluable.

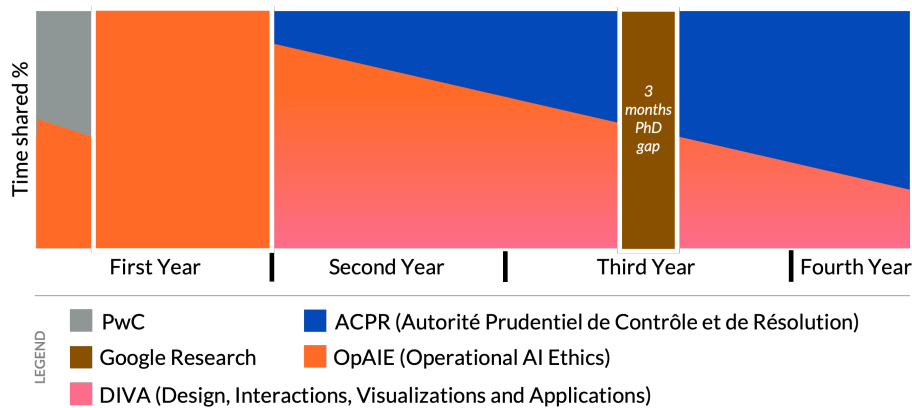


Figure 1: Distribution of my time between different work environments during my PhD, inspired from [Huron, 2014].

My PhD funding comes from the Explainability for Anti-Money Laundering and Counter Terrorism Financing (AML-CFT)² research chair, sponsored by the Agence Nationale de la Recherche (ANR) through the grant ANR-20-CHIA-0023-01 and several private partners, including PricewaterhouseCoopers (PwC), an international consulting firm, Dataiku, a French AI services provider startup, the Crédit Agricole, a large French bank, and the ACPR³, the French Regulatory Authority for Financial Services.

² <https://xai4aml.org/>
³ Acronym for "Prudential Control and Resolution Authority".

As a result, I had various collaboration opportunities with these partners to carry out applied research.

At the beginning of my PhD, I worked briefly with PwC on a survey about AI use in AML-CFT. It enabled me to gain a better understanding of the AML-CFT context. However, the pace of consulting and of research often proved incompatible. Therefore, I focused my first year on academic research, with little industrial collaboration.

The most fruitful collaboration I benefited from was with the Fintech-Innovation team of the ACPR, thanks to Olivier Fliche and Christine Saidani, starting from my second year. This collaboration gave me access to industry expertise and guidance. As a result, I was able to find real-world applications and research questions for XAI in finance that I would not have been able to find alone. More than that, it has provided me with a thriving working environment: welcoming and inspiring colleagues, a second office (close to my home), and a sense of belonging to the Fintech-Innovation team.

Between September and December 2022, I took a break from my PhD to do a research internship at Google's "People and AI Research" team in Toronto. I wrote an interactive article about Saliency Maps, a set of techniques to understand how computer vision models work. It was a fun and rich adventure.

Thanks to

I extend my heartfelt appreciation to my PhD supervisor, Winston Maxwell, whose unwavering support has been instrumental in guiding me through the intricate journey of my doctoral research. You have not only provided invaluable insights into the interdisciplinary and legal aspects of my research but your mentorship has been a beacon of inspiration, making me grow both professionally and personally. I would also like to express my gratitude to James Eagan, my co-supervisor, who joined my advisory team midway through my PhD journey. You brought a fresh perspective to my research, offering invaluable insights to help me better anchor in the discipline of HCI. I am deeply grateful for your thoughtful and friendly guidance.

Thank you also to the OpAIE and DIVA team members for their friendly support. I want to thank Rafik Belloum, Joshua Brand, Mélanie Gornet, Simon Delarue, Tiphaine Viard, Elise Bonnail and Xavier Vamparys for being great colleagues and friends to work with.

A special thanks goes to Olivier Fliche and Christine Saidani, who afforded me the unique opportunity to collaborate with the Autorité de Contrôle Prudentiel et de Résolution (ACPR). Olivier, your expert insights and guidance have been a cornerstone in this dissertation. Thank you for your continued support. Christine, your expertise but also your friendship and shared interest in pottery made my experience particularly enriching and warm. Thanks also to David Bounie for making this fruitful collaboration possible.



Figure 2: This saliency map highlights (in white) the pixels that cause an AI model to recognise this cat as "cat". Find out more in this article I wrote during my Google internship:

<https://pair.withgoogle.com/explorables/saliency/>

Thank you also to the entire FinTech-SupTech team, Jules, Julien, Laurent, Matthieu, Nicolas, Timothée, Lucas for making me feel so welcome, to Laurent Dupont for his help on the Robex experiment, and to all the participants in my user studies at the ACPR for their cooperation and willingness to share their expertise. Working with such a benevolent and knowledgeable group has been an enriching experience that significantly contributed to the depth of my research.

To my family and friends, your support and understanding have been a constant source of encouragement. Benoît, your support has been a driving force throughout this demanding journey.

Thanks also to Samuel Huron, David Cortés, Jan Gugenheimer and Wendy Mackay for their valuable insights and constructive feedback that have greatly enhanced the rigor and clarity of my dissertation. In addition, I thank Tim Miller for his very relevant and thorough comments in his report, and Fosca Giannetti, Alexandre de Streel and Nadia Boukhe-
lifa for their very helpful and constructive feedback.

Contents

<i>Abstract</i>	1
<i>Résumé</i>	3
<i>Acknowledgments</i>	5
<i>List of Figures</i>	18
<i>List of Tables</i>	19
<i>List of Definitions</i>	20
1 Introduction	21
1.1 Research scope	24
1.1.1 <i>Defining AI — not a walk in the park</i>	24
1.1.2 <i>Towards trustworthy AI — and humans</i>	26
1.1.3 <i>HCI and legal perspectives collide in the human-centric approach</i>	29
1.1.4 <i>Explainability (may) contribute to warranted trust</i>	30
1.1.5 <i>Explainability (may) contribute to lawful AI</i>	31
1.1.6 <i>Research domains</i>	36
1.2 Problem statement	38
1.3 Thesis overview	39
1.4 Research approach	41
1.5 Major findings	44

1.6	<i>Academic publications</i>	45
2	Background	47
2.1	<i>A historical perspective on explainability</i>	47
2.2	<i>Explainability in Computer Science: the toolbox</i>	51
2.2.1	<i>The wide range of explainability methods</i>	51
2.2.2	<i>The technical challenges in generating explanations</i>	55
2.3	<i>Explainability in the Social Sciences: the foundations</i>	57
2.3.1	<i>The role of explanations</i>	57
2.3.2	<i>The explanation process</i>	58
2.3.3	<i>Explanations are contrastive</i>	60
2.4	<i>Explainability in HCI: user and context first</i>	61
2.4.1	<i>The need for user-centered explainability</i>	61
2.4.2	<i>Different audiences, different goals</i>	62
2.4.3	<i>Understanding user needs in context</i>	63
2.4.4	<i>Designing explainability systems</i>	65
2.4.5	<i>Evaluating explainability systems</i>	66
2.5	<i>Explainability in Law: dreaming in color?</i>	69
2.5.1	<i>Legal requirements for algorithmic explainability</i>	69
2.5.2	<i>Legal objectives for explainability</i>	72
2.5.3	<i>Is explainability the best disinfectant?</i>	74

PART I CALIBRATING TRUST IN EXPLAINABLE AI: COMMON PITFALLS AND THE PROMISE OF INTERACTIVITY

3	<i>Trust, overtrust, distrust in explainable AI: a cognitive approach</i>	81
3.1	<i>Motivation and research questions</i>	82
3.2	Background	83
3.2.1	<i>Trust in automation</i>	83
3.2.2	<i>Trust in automation by AI systems</i>	84

3.2.3	<i>Explanations are biased and (maybe) biasing</i>	86
3.3	Methodology	89
3.3.1	<i>Review type</i>	89
3.3.2	<i>Corpus creation</i>	89
3.4	Results	92
3.4.1	<i>Overview</i>	93
3.4.2	<i>Cognitive mechanisms explanations should adapt to</i>	94
3.4.3	<i>When explainable AI leads to overtrust</i>	97
3.4.4	<i>When explainable AI leads to distrust</i>	99
3.4.5	<i>When explainable AI is misused</i>	100
3.4.6	<i>When explainable AI corrects false beliefs</i>	100
3.4.7	<i>When explanations are misevaluated</i>	102
3.4.8	<i>Explanations tend to increase unwarranted trust</i>	105
3.4.9	<i>Important factors for appropriate trust: a Bayesian approach</i>	105
3.5	Discussion	107
3.5.1	<i>Take into account cognitive mechanisms and biases in the design of explainable AI</i>	107
3.5.2	<i>Clarify the normal vs. problematic biases with empirical and normative work</i>	108
3.5.3	<i>Detail taxonomies of user groups with cognitive factors</i>	109
3.5.4	<i>Improve our perception of users' reactions to XAI</i>	109
3.5.5	<i>Focus on strategies beyond XAI: contextualization, training, timing, cognitive forcing...</i>	109
3.5.6	<i>Give arguments against the prediction</i>	110
3.6	Limitations	111
3.7	Conclusion	111
4	Towards "human-like" explanations: the promise of interactivity	113
4.1	Motivation and research Questions	114
4.2	Background	116
4.2.1	<i>Interactivity in HCI</i>	116
4.2.2	<i>Interactivity in Explainability</i>	117
4.2.3	<i>Interactivity for learning and sensemaking</i>	118

4.3 Methodology	119
4.3.1 <i>Review type</i>	119
4.3.2 <i>Corpus creation</i>	120
4.3.3 <i>Analysis and coding book</i>	122
4.4 Results	125
4.4.1 <i>Interactivity types in explainability: Select, Mutate, Dialogue with</i>	125
4.4.2 <i>Context, content and form of interactive explanations</i>	130
4.4.3 <i>Evaluating interactive explanations</i>	135
4.4.4 <i>Interactive explanations increase trust, but not necessarily overtrust</i>	138
4.4.5 <i>Interactive explanations are useful, but not easy to use</i>	140
4.5 Discussion	142
4.5.1 <i>Interactivity calls for meta explanations</i>	142
4.5.2 <i>Are dialogic explanations really the grail?</i>	143
4.6 Limitations	144
4.7 Conclusion	145

PART II COMPLYING WITH REGULATION USING HUMAN-CENTRIC EXPLAINABLE AI: TWO CASE STUDIES IN FINANCE

5 Empowering customers of robo-advisors with explainability	151
5.1 Motivation and research questions	153
5.2 Background	154
5.2.1 <i>Mitigating overreliance issues for non experts</i>	154
5.2.2 <i>Designing visualisations of AI explanations for non-expert users</i>	155
5.2.3 <i>Context: life-insurance distribution with "robo-advisors"</i>	155
5.3 Study 1 Methodology: a market-driven co-design approach	158
5.3.1 <i>System design: Robex, the robo-advisor</i>	158
5.3.2 <i>Explanation prototype</i>	163
5.3.3 <i>Co-design sessions and analysis</i>	163

5.4 Study 1 Results	166
5.4.1 <i>Understanding explanation needs from two perspectives</i>	166
5.4.2 <i>Redesign principles drawn from the co-design sessions</i>	166
5.5 Study 2 Methodology: A deception-based between-subjects experiment	170
5.5.1 <i>A 2x4 factorial design</i>	170
5.5.2 <i>Survey procedure and analysis</i>	176
5.6 Study 2 Results	178
5.6.1 <i>Explanations do not help to better calibrate trust</i>	178
5.6.2 <i>Dialogic explanations increase subjective trust</i>	179
5.6.3 <i>Dialogic or graphical explanations do not improve user understanding</i>	179
5.6.4 <i>Explanations do not affect cognitive load and user engagement</i>	179
5.6.5 <i>Higher levels of education reduce overreliance</i>	180
5.7 Discussion	181
5.7.1 <i>Dialogic vs. Graphical explanations</i>	181
5.7.2 <i>Legal requirements for feature-based explanations</i>	181
5.8 Limitations	182
5.9 Conclusion	183
6 Understanding the supervisors' needs for explainable AI in financial crime detection	185
6.1 Motivation and research questions	186
6.2 Background	188
6.2.1 <i>HCI work on eliciting user explainability needs</i>	188
6.2.2 <i>Designing AI justifications for compliance</i>	188
6.2.3 <i>Auditing AI systems</i>	189
6.2.4 <i>The AML-CFT context</i>	189
6.3 Methods	192
6.3.1 <i>Scenario-based semi-structured workshops</i>	193
6.3.2 <i>Empirical legal research</i>	197

6.4 Results	199
6.4.1 <i>Socio-techno-legal context and auditing approaches of supervisors in AML-CFT</i>	199
6.4.2 <i>What provisions in AML-CFT laws does AI opacity conflict with?</i>	203
6.4.3 <i>Supervisors' needs for model justifiability in AML-CFT</i>	205
6.5 Discussion	208
6.5.1 <i>The role of explanations for justifications</i>	208
6.5.2 <i>Considering the limits of explanations</i>	209
6.5.3 <i>Supporting model performance measurement and testing</i>	210
6.6 Limitations	212
6.7 Conclusion	212
7 Discussion	217
7.1 Research contributions	217
7.2 The potential of explanations to manipulate decision-subjects' trust	220
7.2.1 <i>The Self-governance fallacy</i>	220
7.2.2 <i>The dark pattern potential of explanations</i>	221
7.2.3 <i>Safeguards against user manipulation for critical online decisions</i>	221
7.3 Human-centric directions for improved customer empowerment	222
7.3.1 <i>Thinking beyond information access</i>	222
7.3.2 <i>Tailoring explanations to relevant user communities</i>	223
7.3.3 <i>Stimulating skepticism</i>	224
7.3.4 <i>Presenting a selected range of options</i>	225
7.3.5 <i>Fostering user engagement, curiosity and learning</i>	225
7.4 The human-centric way forward for explainability in a highly regulated environment	229
7.4.1 <i>AML-CFT illustrates the tension of using AI in a highly regulated environment</i>	229
7.4.2 <i>Explainability is incomplete and uncertain</i>	230
7.4.3 <i>Human-centric explainability alleviates some of the regulatory tension of black-box AI</i>	231
7.5 Peripheral observations	233
7.5.1 <i>Why the financial sector is interesting for other highly-regulated industries</i>	233

7.5.2 Principles for dealing with interdisciplinarity	234
7.5.3 On explainability for LLMs	235
7.6 General conclusion	238
Appendix	241
A1. List of cognitive patterns when interpreting explainable AI	241
B1. Co-design Study Questionnaire	243
B2. The Robex recommendation system	244
C1. Workshop guide	246
C2. Compliance assessment	248
Bibliography	251

List of Figures

1	<i>Distribution of my time between different work environments during my PhD, inspired from [Huron, 2014].</i>	5
2	<i>This saliency map highlights (in white) the pixels that cause an AI model to recognise this cat as "cat". Find out more in this article I wrote during my Google internship:</i>	6
1.1	<i>John MacCarthy plays chess against a computer in 1967 at Stanford.</i>	24
1.2	<i>AI subdisciplines and their relations from [High-Level Expert Group on AI (HLEG), 2018].</i>	24
1.3	<i>A Geographical Perspective on Explainability. Comparison of keyword searches for "explainability" and "interpretability" on Google from 2004 to present. Shows that China only uses "interpretability", while Israel and Viet-Nam only use "explainability".</i>	30
1.4	<i>Visual representation of the core notions used in this dissertation. We focus on one of the three pillars defined by the HLEG of trustworthy AI: lawful AI. Specifically, we examine the role of explanations to support justifications of AI systems with respect to regulations or regulatory objectives.</i>	32
1.5	<i>The concept of warranted trust and the trust relationships explored in this dissertation. We investigate whether explanations can enhance warranted trust between an individual subject to an AI decision and the AI system, as well as whether explanations can contribute to the development of justifications that support warranted trust between a regulator and the AI system of a regulatee.</i>	35
1.6	<i>Domain scope</i>	36
1.7	<i>Topic network of the FAT and Interpretable ML community in [Abdul et al., 2018].</i>	37
1.8	<i>Overview of the work presented in this dissertation through a modified version of the triangulation framework of Mackay and Fayard [1997], inspired from [Huron, 2014]</i>	41
2.1	<i>A Historical Perspective on Explainability. The bar plot (in red) shows the evolution of the number of academic contributions on XAI. The bubble chart on top displays the number of citations—represented by size and y-axis—of the most influential papers in XAI.</i>	49
2.2	<i>Distribution of contributions in explainable AI accross disciplines. This graph is based on a corpus of 5756 articles published from 2015 to present, extracted from searching "explainab*" in the article title in the Scopus Database.</i>	50
2.3	<i>Categorization of explainable AI methods along four dimensions inspired by Nauta et al. [2023] and Barredo Arrieta et al. [2020].</i>	51
2.4	<i>Illustrative examples of feature-based explanations for different data types (image, tabular and text data) with input saliency [Alammar, 2021, Unruh and Robinson, 2020].</i>	52
2.5	<i>Illustration of the gradient-based method to identify "salient" pixels.</i>	53
2.6	<i>Figure 1 in [Tomsett et al., 2018] identifies the different stakeholders in a machine learning ecosystem. "Direction of arrow indicates direction of interaction."</i>	62
2.7	<i>The four reasons motivating the need for explainable AI presented in [Adadi and Berrada, 2018].</i>	63

2.8	<i>Examples of visual explanations for different AI models a) Hybrid visual and textual explanations for the estimation of the reading time of an article [Szymanski et al., 2021], b) Influence of features on loan default risk [Chromik et al., 2021], c) Multiple explanations for house price forecasts [Hohman et al., 2019]), d) Example-based explanation for drawing recognition [Cai et al., 2019].</i>	66
2.9	<i>The 12 Explanation quality properties proposed by [Nauta et al., 2023].</i>	67
3.1	<i>PRISMA flow diagram [Moher et al., 2009] on how the final corpus was curated (n = 38).</i>	90
3.2	<i>The distribution of the corpus across disciplines.</i>	92
3.3	<i>Summary of the cognitive constraints, biases and mitigation strategies discussed in the papers included in our corpus (n=38).</i>	93
3.4	<i>The 38 papers in the corpus and a rough indication of whether the paper reports on over- or distrust effects of explanations, on the misuse of explanations, or on other explanation-related phenomena.</i>	103
4.1	<i>Summary of the role of explanations, the process by which we construct and present explanations and the biases involved in explanations.</i>	114
4.2	<i>Illustrative example of interactive explanation: "Conversational XAI" enables users to interact with users through natural language.</i>	114
4.3	<i>Illustrative example of interactive, rule-based explanation where users can create and modify rules.</i>	115
4.4	<i>PRISMA flow diagram adapted from Page et al. [2021] giving an overview of the PRISMA 2020 survey guidelines, used for the search and selection phases of our scoping review. . .</i>	119
4.5	<i>Example of the clarify interaction taken from [Anik and Bunt, 2021].</i>	126
4.6	<i>Examples of the arrange interaction taken from [Hohman et al., 2019] (top) and [Cheng et al., 2021] (bottom).</i>	127
4.7	<i>Examples of the filter/focus interaction taken from [Hohman et al., 2019] (top) and [Ming et al., 2019] (bottom).</i>	127
4.8	<i>Examples of the reconfigure interaction taken from [Ming et al., 2019] (top) and [Collaris and van Wijk, 2020] (bottom).</i>	127
4.9	<i>Examples of the simulate interaction taken from [Ross et al., 2021] (top) and [Cheng et al., 2019] (bottom).</i>	128
4.10	<i>Example of the compare interaction taken from [Hohman et al., 2019].</i>	128
4.11	<i>Example of the progress interaction taken from [Melsión et al., 2021].</i>	129
4.12	<i>Examples of the answer interaction taken from [Melsión et al., 2021] (top) and [Guo et al., 2022] (bottom).</i>	129
4.13	<i>Example of the ask interaction taken from [Melsión et al., 2021].</i>	129
4.14	<i>"Interactive XAI helps users. . ."</i> <i>Illustration of the taxonomy of interaction in explainability with screenshots from the corpus.</i>	130
4.15	<i>Left: Frequency of the interaction categories used in the corpus and frequency of their combinations ; Middle: Percentage of studies using an explanation representation per interaction category; Right: Percentage of studies focusing on a type of user question per interaction category/</i>	132
4.16	<i>The first part of the concept matrix [Webster and Watson, 2002], reporting the explanation context and content. The design of this concept matrix was inspired from [Bae et al., 2022].</i>	133
4.17	<i>The second part of the concept matrix, reporting the explanation communication and evaluation.</i>	134

4.19	<i>Left: Count of the positive, negative and neutral quantitative evaluations of interactive explanations compared to static ones, against various user-based metrics, based on 9 different studies. Right: Count of the different evaluation outcomes in the empirical studies comparing interactive explanations with no explanation as a baseline, extracted from 13 different papers in the corpus.</i>	139
5.1	<i>Fictional life-insurance plans proposed by Robex, the explainable robo-advisor developed for this study</i>	159
5.2	<i>Screenshot of the Robex interface, showing the profiling questionnaire stage at the start of the user journey. Translated from French to English.</i>	160
5.3	<i>Screenshot of the Robex interface, showing the recommendation stage. As required by law, a summary of the user's profile is displayed first, followed by a life-insurance contract proposal with details. The explanation is presented on the same page, just after the proposal.</i>	160
5.4	<i>Screenshot of the Robex interface, showing the answers it provided for test questions on participants' financial knowledge.</i>	162
5.5	<i>Screenshot of the feature-based explanation prototype for Robex. In original language (French). Individual factors that decrease investment risk are shown on the left in descending order of importance and factors increasing investment risk are on the right. . . .</i>	164
5.6	<i>Explanation interfaces for each of the condition A "Graphical-static": users see a graphical summary of how their characteristics impact the risk of the proposal. Translated from French to English.</i>	171
5.7	<i>Explanation interfaces for each of the condition B "Graphical-mutable": users first see the graphical-static interface and then a pop-up message indicates they can change some of their characteristic. Translated from French to English.</i>	172
5.8	<i>Explanation interfaces for each of the condition C "Dialogic": the same information provided in the interfaces A and B) is delivered through "sms-like" textual messages. Some graphics are added to facilitate the visualisation of the risk and of the variables decreasing and increasing the risk of the proposal. Translated from French to English.</i>	173
5.9	<i>Explanation interfaces examples for an incorrect recommendation for each of the three conditions: A' "Graphical-static"; B' "Graphical-mutable"; C' "Dialogic". The correct user profile in this case would have been "Secure", but the skewed Robex algorithm outputs "Dynamo". Only A' is translated from French to English, the rest are in original language.</i>	174
5.10	<i>The workflow of our quantitative experiments. The profiling questionnaire is used to produce a personalized recommendation of a life-insurance contract. Clients can review the recommendation, the explanation and then decide to follow the recommendation or not.</i>	176
5.11	<i>Results for Study 2. Vertical lines represent the 95% confidence interval. Asterisks and dots indicate the statistical significance of the results: *** $p\text{-value} \leq 0.001$, ** $p\text{-value} \leq 0.01$, * $p\text{-value} \leq 0.05$, • $p\text{-value} \leq 0.07$, "ns" non significant.</i>	178
5.12	<i>Effects of education on reliance, understanding of the recommendation and of the explanation.</i>	180
6.1	<i>Scenarios used during the workshops with supervisors, with a description of the two use cases of AI in AML-CFT, and two examples of alerts that were generated or closed by the AI-enhanced systems. Only one of these case studies was presented in each workshop. . .</i>	194
6.2	<i>Conceptual justifications shown for the scenario 2 and its example alert. Conceptual justifications for the scenario 1 followed the same format.</i>	195
6.3	<i>Summary of the workshops, with socio-techno-legal context of supervisors, supervisors' questions on AI, AI auditing approaches ideas and ideas for justifications and explanations.</i>	199
6.4	<i>Flow diagram of the supervisor's control procedures in AML-CFT</i>	202

7.1 *Explanation interface to engage users cognitively and stimulate their curiosity. First, a brief explanation of Robex is given: a); second, the user answers several multiple choice questions that lead them to question the impact of some features: b) and c); third, the full graphical explanation is given. 228*

List of Tables

2.1	<i>The different classifications of audiences, goals, explanation content, explanation timing and contexts presented in the XAI literature.</i>	64
3.1	<i>Coding book used for the analysis of the corpus.</i>	92
4.1	<i>Codebook used to retrieve information from the corpus with four dimensions: [explanation] context, content, communication and evaluation, their corresponding sub-dimension and reference from which codes were inspired from.</i>	123
4.2	<i>Two-level taxonomy of interactivity techniques in XAI, including a first level reflecting the type of support interaction techniques provide to the cognitive process of explaining, a second task-oriented level, and corresponding definitions.</i>	126
5.1	<i>Question used in the Robex's profiling questionnaire for measuring users' personal characteristics (translated from French to English).</i>	161
5.2	<i>Main themes emerging from the content analysis of supervisors and end-users interviews, with corresponding lexical field and citations.</i>	167
5.3	<i>Question used for measuring different metrics with Cronbach alphas (translated from French to English).</i>	175
6.1	<i>Description of role, experience, familiarity with AI of participants in the study.</i>	213
6.2	<i>Data used for the empirical legal research</i>	214
6.3	<i>Summary of supervisors' needs for model justifiability, corresponding description, model concerned and developer of justifications/explanations, and justification and explanation design ideas that emerged during the workshops.</i>	215

List of Important Definitions

1.1	<i>Artificial Intelligence (McCarthy and Minsky, 1956)</i>	24
1.2	<i>AI system (OECD, 2023)</i>	25
1.3	<i>Trustworthy AI</i>	26
1.4	<i>Trust</i>	27
1.5	<i>Warranted trust</i>	28
1.6	<i>Trust calibration</i>	28
1.7	<i>Overtrust and Distrust</i>	28
1.8	<i>Overreliance or Underreliance</i>	29
1.9	<i>Human-centric AI</i>	29
1.10	<i>Explanation</i>	30
1.11	<i>Explainability</i>	30
1.12	<i>Explainable AI (XAI)</i>	30
1.13	<i>Interpretable AI</i>	31
1.14	<i>Regulation</i>	32
1.15	<i>Audit, auditability</i>	33
1.16	<i>Accountability</i>	33
1.17	<i>Justification</i>	34
3.1	<i>Complacency</i>	83
3.2	<i>Automation bias</i>	83
3.3	<i>Cognitive biases</i>	87
4.1	<i>Perceived usability</i>	136
4.2	<i>Perceived usefulness</i>	136
7.1	<i>Dark patterns</i>	221
7.2	<i>User engagement</i>	225
7.3	<i>Curiosity</i>	226
7.4	<i>RegTech</i>	234

Chapter 1

Introduction

“High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system’s output and use it appropriately. An appropriate type and degree of transparency shall be ensured, with a view to achieving compliance with the relevant obligations of the user and of the provider.”

Proposal for the AI Act, April, 21st, 2021

AI IS HYPE TODAY. 2023 was Generative AI’s breakout year, with ChatGPT and Midjourney¹ generating significant excitement around perfectly credible presidential speeches produced in a few seconds or videos of teddy bears skating. However, AI’s large scope of benefits, from personalized movie recommendations to the detection of cancerous lesions in medical imaging, comes with risks. Public and expert opinions have expressed concerns about AI taking over human jobs, people gradually losing skills, or privacy and fundamental rights being violated by AI decision systems [Cui, 2023, Zhang, 2021]. Notably, the use of AI in automated settings has fueled concerns about AI replacing humans, and the need for keeping humans in control for important decisions. In a recent survey, [Tyson and Kikuchi, 2023] highlighted that Americans’ concern about AI in daily life outweighed excitement.

Many concerns arise from the **complexity and opacity** of some AI models, and more specifically deep learning². While we know the mathematical operations that occur in perceptrons, units of neural networks inspired by brain neurons [Cox and Dean, 2014], we do not understand why, when put together, they result in the behavior we observe [Anthropic, 2023]. The scale of the data on which these models are trained, and the massive number of parameters that compose them³ makes them unintelligible to humans. Like the human brain, we have a good understanding of its component units, such as synapses, and how they communicate with each other, but we cannot fully explain the results they produce [Anthropic, 2023]. Sophisticated machine learning models, especially generative ones, are often considered as “black-boxes”. They can

¹ Models like ChatGPT or MidJourney, which create text or images from prompts, are called “generative AI”. <https://chat.openai.com/>
Accessed January 2024.

² Deep learning is a subset of machine learning which involves neural networks with multiple hidden layers.

³ GPT-4, for example, has 1.7 trillion parameters.

provide very accurate predictions, but it is unclear how they arrive at those conclusions.

The emergence of deep learning models in 2012 [Krizhevsky et al., 2012, LeCun et al., 2015] and more recently, transformers [Vaswani et al., 2017] and generative AI, has brought us in what Melanie Mitchell [2021] describes as an "AI spring", a period of massive investment and optimism in AI. This "race to AI" has led to a "**race to regulation**" [Smuha, 2021]. Regulatory efforts to prevent the harmful effects of AI systems have multiplied in recent years, the results of which are only now starting to emerge. In Europe, the proposal for the regulation of AI in the European Union (the "AI Act") [European Commission, 2021], which sets out requirements for AI applications considered as "high risk" will require thorough certification mechanisms for machine learning systems considered as "high risk". China has also adopted a set of regulations following its "Next Generation AI Development Plan" [Zheng and Zhang, 2023] in 2017⁴. In the United States, the most recent federal regulatory effort consists of the White House executive order on AI [The White House, 2023], laying down principles for responsible AI⁵. In parallel, questions arise about the compliance of AI systems with existing regulatory frameworks, particularly in highly regulated areas with well-established norms [Mittelstadt et al., 2019].

A key objective of regulation is to protect end users and citizens from various detrimental consequences such as being deceived, being discriminated against, or suffering from algorithmic errors. As a result, many of the AI policies detailed above present transparency as a central theme. Some existing, sector-specific, regulatory frameworks already impose obligations to explain an algorithmic prediction to the end user. This is the case, for example, in the context of protecting customers of online life insurance recommendation systems. In other situations, the use of machine learning models in regulated environments requires explanations addressed to regulators in charge of verifying the compliance of the system. Shedding light on the complex inner workings of AI models has been the subject of an entire field of research called *explainability (XAI)*, which has gained considerable interest over the last five years. In particular, the research and policy communities have become increasingly aware of the importance of "human-centric" design of AI explanations. However, little attention has been paid so far to the human-centric design of explanations in view of demonstrating compliance with applicable regulation and ensure "**lawful AI**" [High-Level Expert Group on AI (HLEG), 2018].

In this thesis, we show through literature reviews and experiments in the context of life-insurance online distribution that AI explanations can have the paradoxical effect of increasing user trust, including unwarranted trust. Instead of empowering them, explanations can make non-expert users more vulnerable. This may undermine the regulatory objectives to inform and "enlighten" customers about the AI-based decisions being made about them. We also identify the different ways in

⁴This plan includes the 2022 "Administrative Provisions on Algorithm Recommendation" [Zheng and Zhang, 2023] and the world's first Generative AI Regulation published in August 2023.

⁵Additionally, ten states have regulated the use of AI, including hiring and profiling algorithms, as part of broader consumer privacy laws [Katrina Zhu, 2023].

which explanations can lead AI users to overtrust, distrust, or misunderstand the system. Additionally, we investigate the effects of more interactive "human-like" explanations that could avoid the identified pitfalls. We argue that better efforts can be made to create more effective AI explanations through the human-centric approach, by supporting user engagement, curiosity and learning.

We also discuss how explainability can contribute to building justifiable trust of AI stakeholders, including regulators, in the context of anti-money laundering and countering terrorism financing (AML-CFT). The success of explainability for regulators will depend on taking a human-centred approach designed to avoid human biases and adapt to the socio-technical features of this context. We highlight that current explainability methods have severe limitations and may contribute to an unjustified sense of certainty about AI systems' behavior. However, human-centric explainability can still help alleviate the tensions created by the use of black-box AI systems in AML-CFT by contributing to justifiability and accountability.

1.1 Research scope

This section outlines some key terms and ideas necessary to understand the scope and motivation of this dissertation. It then details the research domains in which it falls.

1.1.1 Defining AI — not a walk in the park

AI is a broad church [Boden, 1996]. There may exist as many definitions as there are people who use it [Smuha, 2021]. One working, illustrative definition was given by John McCarthy of MIT and Marvin Minsky of Carnegie-Mellon in the context of the 1956 Dartmouth College. They defined AI as:

Definition

Artificial Intelligence (McCarthy and Minsky, 1956). *The construction of computer programs that engage in tasks that are currently more satisfactorily performed by human beings because they require high-level mental processes such as: perceptual learning, memory organization and critical reasoning” [Council of Europe, 2023].*

For example, playing chess, driving a car, translating, are examples of tasks that require complex acquisition and reasoning processes including vision, spatial awareness, judgment [Surden, 2019], and which AI was being programmed to achieve.

The decades between 1950 and 1990 were the years of fundamental advances in neural networks⁶ and "symbolic artificial intelligence" which was based on knowledge and reasoning representation. Expert systems built in the 1980s mirrored human logic in their "inference engine" and marked the golden age of symbolic AI. In the 2010s, access to massive amounts of data and the development of powerful processors made it possible to fully exploit the ideas previously developed on neural networks [Council of Europe, 2023]. Instead of coding human-driven logical rules in computers, the neural network or machine learning approach relied on letting systems discover rules by themselves in the data.

It is generally considered that four types of machine learning exist: supervised, semi-supervised, unsupervised, and reinforced learning. Following Ghahramani [2004]’s definitions, *"in supervised learning the machine is given a sequence of desired outputs y_1, y_2, \dots , and the goal of the machine is to learn to produce the correct output given a new input."* In unsupervised learning, however, *"the machine simply receives inputs x_1, x_2, \dots , but [does not] obtain supervised target outputs"*. For instance, clustering is a common unsupervised learning technique where the machine finds groups of data that share similarities. In semi-supervised learning, the machine generates its own targets y_1, y_2, \dots to "supervise itself". In reinforcement learning the machine gets rewards whenever its forecast or behavior is correct.

From this historical perspective, artificial intelligence encompasses all systems designed to imitate, match or surpass problem solving skills



Figure 1.1: John McCarthy plays chess against a computer in 1967 at Stanford.

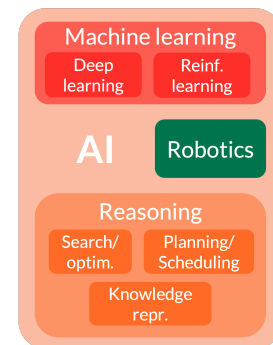


Figure 1.2: AI subdisciplines and their relations from [High-Level Expert Group on AI (HLEG), 2018].

⁶ Neural networks were invented much earlier than the AI boom of 2012. For example, the idea of the ReLU function was presented in 1969 by Fukushima, backpropagation was invented in 1970 by Linnainmaa, LSTM were introduced in 1995 by Hochreiter and Schmidhuber, etc. [Hochreiter and Schmidhuber, 1997, Müller et al., 1995]

of the human brain, from symbolic AI, to machine and deep learning or robotic systems. While illustrative, this definition carries the risk of mistaking AI for actually intelligent, thinking, or even sentient, machines [Mitchell, 2021, Surden, 2019]. Surden [2019] argues that it is essential to understand what AI is not, emphasising that the computational processes it employs are nothing like human thinking: "AI systems are often able to produce useful, intelligent results without intelligence".

AI must therefore be defined differently from the objective of matching or surpassing human intelligence, which is either evasive, speculative or even misleading. Recent attempts at aligning AI policy have provided alternative definitions that offer a functional, rather than intentional description, based on the capabilities that AI systems demonstrate. In 2019, the OECD proposed a definition for AI systems in the "Recommendation of the Council on Artificial Intelligence", that was adopted by 38 countries. The definition was amended on November, 8th, 2023:

Definition

AI system (OECD, 2023). *A machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment [OECD, 2019].*

The AI Act [European Commission, 2021] considers a similar definition of AI systems in Article 3 "Definitions"⁷.

However, defining AI in legal terms has proven difficult, giving rise to wide-ranging political discussions and academic debates. Some scholars have argued that agreeing on a single definition of AI was unfeasible [Reed, 2018], or even undesirable [Schuett, 2023]. Schuett [2023] contends that policy makers should not use the term AI, which does not comply with common requirements for legal definitions. These requirements stem from general legal principles of democratic countries, such as the principle of proportionality, effectiveness, legal certainty or the vagueness doctrine. With regard to these principles, Schuett argues that artificial intelligence is too vague, over-inclusive, unpractical, imprecise and unintelligible of a term to be used as a legal definition.

⁷It also draws on the definition proposed by the High-Level Expert Group in 2018 [High-Level Expert Group on AI (HLEG), 2018].

Aware of all of these difficulties to delineate the scope of AI, this dissertation nonetheless focuses on algorithmic systems that fit the second definition provided above. In the first part of the dissertation, we will be particularly interested in how these systems "influence" their environment, and more specifically human operators, when used as decision aids. In the second part, we will narrow our focus to AI systems used in finance. The first use case is an expert system providing recommendations for life-insurance contracts. The second use case explores different types of machine learning systems, supervised and unsupervised, to detect money laundering and terrorism financing.

1.1.2 Towards trustworthy AI — and humans

Against a backdrop of surging investments and competition in AI, research has shown that AI could cause harms, intended or not⁸, such as discrimination, wrongful arrests, spreading of fake-news, defamatory deep-fakes, among others [Acemoglu, 2021]. In response to AI-specific risks, a multitude of ethical principles for AI have emerged. A notable success in aligning different stakeholders at scale was achieved with the OECD principles developed in 2019 and endorsed by 46 countries [OECD, 2019]. The OECD proposed ten principles for AI, which represent a set of priorities to reflect democratic values in AI policies, such as protecting human rights, equity, or establishing stakeholder accountability. A similar early attempt at characterizing desirable AI properties comes from the 2019 Guidelines for Trustworthy AI by the HLEG [High-Level Expert Group on AI (HLEG), 2019]. The guidelines propose seven key requirements that AI systems should meet to be considered trustworthy: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, societal and environmental well-being, and accountability. The guidelines were influential in the drafting of the AI Act [European Commission, 2021]. Other initiatives include the trustworthiness framework for AI proposed by the International Organization for Standardization (ISO) [International Organization for Standardization (ISO), 2022] or the National Institute of Standards and Technology's (NIST) Method for Evaluating User Trust in AI system [NIST, 2023] developed in the U.S. The aforementioned efforts are among the most influential ones, but many other frameworks, ethical guidelines, principles for AI have been proposed by either international, governmental, or private organizations [Kaur et al., 2022, Jobin et al., 2019].

⁸ AI harms studied in such research are mainly not human-intended, however some are direct consequences of poor AI development choices and optimization objectives.

Overall, two umbrella terms have emerged, "Responsible" or "Trustworthy" AI, to embody the ethical and safe use of AI. The former was used mainly by private organisations, and possibly comes from the Corporate Social Responsibility (CSR) culture where the notion of responsibility and accountability are predominant.

The term *trustworthy AI* has emerged as a comprehensive objective for AI systems. It was promoted by the EU strategy for AI [European Commission, 2023] in 2017, the OECD principles, the ISO and NIST frameworks, among others, and places trustworthiness as a higher, ultimate value. The High-Level Expert Group on AI give three conditions for AI systems to be trustworthy: they should be lawful, ethical and robust (*cf.* Figure 1.4) [High-Level Expert Group on AI (HLEG), 2019]. According to Kaur et al. [2022]'s review:

Definition

Trustworthy AI. is a framework to ensure that a system is worthy of being trusted based on the evidence concerning its stated requirements. It makes sure that the users' and stakeholders' expectations are met in a verifiable way [Kaur et al., 2022].

The HLEG's decision to concentrate on the concept of trust is understandable. Trust is pillar of our society and lives. It determines our interactions with people, institutions, organizations and machines. Many distinct conceptual visions of trust have been proposed through the lenses of philosophers, economists or psychologists. In the context of trust in AI, we retain one proposed by Danks [2019] for the remainder of this dissertation, which focuses on the functional value of trust:

Definition

Trust. Condition in which "the user has a reasonable belief that the system (whether human or machine) will behave approximately as intended" [Danks, 2019].

The definition is in line with the one given in [Jacovi et al., 2021]. Following [Lee and See, 2004], Jacovi et al. [2021]'s model of trust also incorporates the dimension of vulnerability: "trust is an attempt to anticipate the impact of behavior under risk". In the case of human-AI trust, the user is vulnerable to the risk of the AI being wrong. Trust makes her believe that the risk is low. This risk-taking element is present in other definitions of trust in the literature [Mayer et al., 1995, Glikson and Woolley, 2020]⁹.

The core value of trust is to enable cooperation [Hardin, 2006]. Trust makes social cooperation easier and even possible [Hardin, 2006]. It also enables cooperation between people and technology [Jacovi et al., 2021, Ferrario and Loi, 2022, Chatila et al., 2021], in part because we often apply the same social norms of interaction with machines as we do with humans [Miller, 2019]. Consequently, trustworthy AI ultimately aims to enable and improve human-AI *cooperation*, or *collaboration* which one objective of human-computer interaction research [Jacobs et al., 2021, Khadpe et al., 2020]. This enriched collaboration between humans and AI systems can also be framed as enhanced decision-making. In critical applications such as healthcare, finance, justice, Chatila et al. [2021] contends that really useful AI systems make it possible for human decision makers to take decisions that are more informed, as free of bias as possible and "ultimately better".

The concept of trustworthy AI is subject, however, to controversy. Critics mainly point to the fact that trustworthy AI and other expressions such as *responsible AI* or *accountable AI* can obscure a necessary, active role for humans, and pose the wrong questions. Joanna Bryson [2018] argues that trust can only be deferred to peers (other human beings), and not to machines [Joanna Bryson, 2018, Smuha, 2021]. As physical and legal entities, humans are the ones who should be "responsible" and "accountable" for AI systems, not AI. Marisa Tschopp [2020] advances that tech companies should ask themselves "How can we be trustworthy?" rather than "How can we we increase trust in AI?". Additionally, some note that the idea of trust, in its philosophical meaning, involves delegating control, in this case to the machine, without the need for supervision [Smuha, 2021, Ferrario et al., 2020]. In fact, due to the opaque nature of machine learning models, AI stakeholders are likely to have to

⁹Mayer et al. [1995] explain that trust implies "taking a meaningful risk while believing in a high chance of positive outcome".

trust an AI system without a complete understanding of its underlying algorithms. Ferrario and Loi [2022] even propose an account of trust as "anti-monitoring", as it goes against the idea of complete comprehensibility and control. As Lee and See [2004] note:

"Trust guides reliance when complexity and unanticipated situations make a complete understanding of the automation impractical".

[Lee and See, 2004], (p. 50).

However, in most situations, it is not desirable that people blindly trust so-called "trustworthy" AI systems. Rather, the goal is to have responsible users able to calibrate their trust by relying on tangible information about the system, provided by measures such as transparency, explainability, safety tests, uncertainty metrics [Kurz et al., 2022], etc. Hardin [2006] states *"I am likely to trust you when you have given some evidence of being trustworthy"*. Jacovi et al. [2021] note that trustworthiness and trust are two entirely disentangled concepts. Trust can exist for an untrustworthy system and vice-versa.

Definition

Warranted trust. *Trust is warranted when it is caused by trustworthiness (to some contract, defined for example by the HLEG's key requirements for trustworthy AI). In the opposite case, it is unwarranted [Jacovi et al., 2021], or misplaced.*

Ferrario et al. [2020] define "paradigmatic trust" as the disposition of individuals to rely on an AI system without monitoring, but having formed beliefs about the system's trustworthiness, through evidence of its reliability¹⁰. Ferrario and Loi [2022] present paradigmatic trust as *justified and warranted* trust. We use hereinafter the terms *warranted* [Jacovi et al., 2021], *justified* [Ferrario and Loi, 2022] or *appropriate* trust as synonyms [Gunning and Aha, 2019].

In this dissertation, we focus on the impact of explanations of AI systems on users' *warranted* trust. This amounts to studying the process of *calibrating* trust.

Definition

Trust calibration. *The process of assigning a level of trust to a system based on its performance, capabilities and behaviour [Culley and Madhavan, 2013].*

Inappropriate trust calibration may lead to *overtrust*, *distrust*, *overreliance* or *underreliance*, i.e. *misplaced* or *inappropriate* trust. We use hereinafter the following definitions for these terms:

Definition

Overtrust and Distrust. *As an excessive or insufficient level of subjective trust. Subjective trust measures the participants' subjective reports of*

¹⁰ This model of trust poses the reduced level of monitoring as an important characteristic of trust. However, we see the reduced levels of monitoring as a consequence of trust and not a defining characteristic of the concept.

trust in the (X)AI system (also called perceived trust) [Bagheri and Jamieson, 2004, Miller, 2022].

Definition

Overreliance or Underreliance. *An excessive or insufficient level of demonstrated trust. Demonstrated trust, or reliance, refers to the propensity of participants to follow and accept the advice or prediction of an (X)AI system [Miller, 2022].*

1.1.3 *HCI and legal perspectives collide in the human-centric approach*

A crucial element of this trust calibration process is human behaviour in the context of receiving AI predictions. Developing trustworthy AI requires understanding the factors and mechanisms in human-AI interactions that contribute to building trust [Jacovi et al., 2021, Danks, 2019]. Work to advance in this direction must therefore adopt a human-centered approach. This endeavour has been characterized as *human-centric AI* in recent literature [Shneiderman, 2020, Maxwell and Dumas, 2023, Bryson and Theodorou, 2019].

Definition

Human-centric AI. *This approach places people and users at the centre of the development of AI [European Commission, 2019]. It promotes the study of AI users in context, to understand their needs. The approach also encompasses the understanding of the cognitive processes that underlie human-AI interactions.*

In recent years, the goal of human-centred AI has become sufficiently clear and consensual for several disciplines to feel concerned, allowing a holistic view of the problem. Specifically, Human-Computer Interaction (HCI) and legal perspectives seem to collide in the human-centric AI approach. HCI is obviously part of the mix of the disciplines involved. Specifically, human-centric AI builds on HCI's long history of user-centred design [Abrás et al., 2004]. However, policy and legal experts have also adopted a human-centric approach to AI, despite law historically being a rather independent academic discipline [Barocas et al., 2020]. For instance, the High-Level Expert Group (HLEG) has fully embraced a human-centric approach [High-Level Expert Group on AI (HLEG), 2019], which is also reflected in the new AI Act regulation¹¹. The regulation also takes into account observations from psychology and Human-Computer Interaction (HCI) literature regarding the impact of human factors on trust, such as the severity of consequences¹².

This thesis falls within the human-centric AI approach. We focus on the calibration of trust between humans and AI systems, through explainability, as a key enabler of meaningful human-AI collaboration. We also complement the human-centric perspective with legal approaches in the case studies presented in Part II.

¹¹ "Rules for AI available in the Union market or otherwise affecting people in the Union should therefore be human centric, so that people can trust that the technology is used in a way that is safe and compliant with the law, including the respect of fundamental rights" in Section 1.1 "Reasons for and objectives of the proposal" [European Commission, 2021]

¹² Recital 38a of the draft proposal

1.1.4 Explainability (may) contribute to warranted trust

Explainability serves as one of the levers to extract information about the behaviour of AI systems [Markus et al., 2021, High-Level Expert Group on AI (HLEG), 2019, Jacovi et al., 2021]. However, there exist terminological nuances and controversies in the definition of explainability [Markus et al., 2021]. Some argue that the term *explainability* and the acronym XAI are reserved for the mathematical methods used to interrogate AI systems and extract insightful information about their inner workings [Herzog, 2022]. Another term, *interpretability* is therefore used to refer to the propensity of an AI system to be contextualized and human-understandable [Broniatowski, 2021]. Additionally, *interpretable AI* usually refers to models that are designed in a way that is simple enough for humans to fully understand them [Rudin, 2019]. As for the term *intelligibility*, it refers to the propensity of an explanation to be human-understandable [Weld and Bansal, 2018].

We can see that the variations between these different terms can be subtle. Moreover, there is no consensus on their definitions in the current literature. For example, some use explainability and interpretability [Markus et al., 2021] interchangeably. Depending on their geographical region, researchers may only use one term and not the other, as shown in Figure 1.3. To clarify the meanings of explainability-related terms, we retain the following definitions for the rest of this dissertation:

Definition

Explanation. Explanations of AI systems are transfers of knowledge about the behavior AI systems [Henin and Le Métayer, 2022, Miller, 2019]. Henin and Le Métayer [2022] state that explanations are "descriptive and intrinsic in the sense that they only depend on the system itself".

Definition

Explainability. Explainability broadly refers to providing explanations of AI systems to relevant stakeholders to scrutinize AI models in their development, implementation, and deployment stages [Herzog, 2022]. It most commonly involves demands for transparency and interpretability of AI systems [Herzog, 2022].

Definition

Explainable AI (XAI). Explainable AI (XAI) is the technical arm that aims to provide explainability. Following Markus et al. [2021] and Gilpin et al. [2018], an AI system is explainable if it is intrinsically interpretable, or if it is complemented with an interpretable and faithful explanation. Interpretability covers aspects related to the intelligible and understandable aspect of explanations by humans. Fidelity captures the capacity of an explanation to provide accurate and truthful accounts of an AI system.

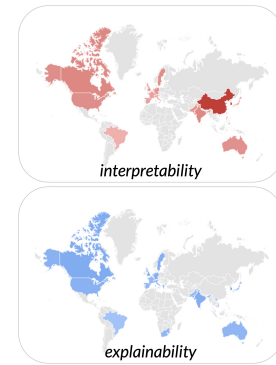


Figure 1.3: A Geographical Perspective on Explainability. Comparison of keyword searches for "explainability" and "interpretability" on Google from 2004 to present. Shows that China only uses "interpretability", while Israel and Viet-Nam only use "explainability".

Definition

Interpretable AI. *A subset of algorithms that are simple enough to be completely understood by design. These include linear and logistic regressions, decision trees and rules, Generative Linear Models and Generative Additive Models [Molnar, 2019].*

Expressions such as the ones we use in this section "explainability contributes to trust", "explainability fosters trust in AI" are common in the recent literature on human-AI collaboration [Ferrario and Loi, 2022]. However, the relationship between explainability and trust is not straightforward and needs to be challenged.

Ferrario and Loi [2022] argue that there exists a causal relationship between the perceived reliability of an AI system, given by reliability indicators, and the perceived trustworthiness of a system. According to the authors, explainability therefore fosters trusts only if it is an indicator of reliability of the AI system. In the context of medical AI, the authors do not believe that explainability can meet this condition, as it does not directly depict how reliable and predictable an algorithm is. They consider that explainability is neither sufficient, nor necessary, to form justified beliefs about the trustworthiness of the AI system. They also note that there is no link between the explainability of a system and the absence of need to monitor it, which for them characterises trust [Ferrario and Loi, 2022]. The authors, however, note that these claims have yet to be demonstrated empirically.

On the contrary, Jacovi et al. [2021] note that explainability enables warranted trust by making possible the observation of the intrinsic reasoning process of the AI system and external symptoms of the model behavior. In other words, explainability is unique in its ability to establish 'intrinsic trust', whereas other mechanisms for establishing calibrated trust rely on 'extrinsic' trust mechanisms. As a result, explainability can foster distrust in a non-trustworthy system and trust in a trustworthy one.

In this dissertation, we aim to further clarify the challenges in enabling warranted trust with explainable AI by reviewing existing practices in the XAI field and conducting empirical studies in the financial sector.

1.1.5 Explainability (may) contribute to lawful AI

As AI enters highly regulated environments, and specific AI regulation emerges, the issue of monitoring compliance of AI systems with existing or new regulations arises. This thesis examines the role of explainability in enabling such controls, and enforcing "lawful"¹³ or compliant AI. More specifically, we look at explanations' role in *justifying* that AI systems are compliant within some set of rules. We consider that, if explainability can contribute to fostering warranted regulator trust and, in specific cases, warranted consumer trust, it participates to making AI "lawful". We examine the challenges in using explainability for demonstrating compliance with specific financial regulations.

¹³ "Lawful AI" is one of the three conditions for trustworthy AI. It is defined as "respecting all existing applicable laws and regulations" [High-Level Expert Group on AI (HLEG), 2018].

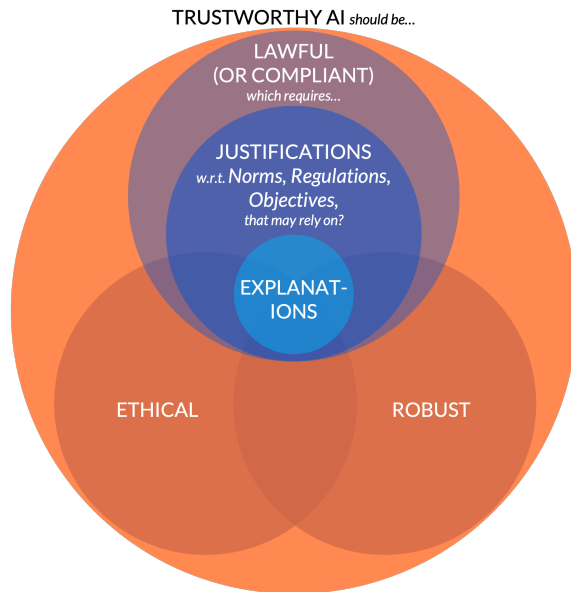


Figure 1.4: Visual representation of the core notions used in this dissertation. We focus on one of the three pillars defined by the HLEG of trustworthy AI: *lawful AI*. Specifically, we examine the role of *explanations* to support *justifications* of AI systems with respect to regulations or regulatory objectives.

Central to the notion of regulation is the power to compel regulated entities to conform to a set of standards. Julia Black proposed the following seminal definition of regulation:

Definition

Regulation. *"The intentional use of authority to affect behaviour of a different party according to a set of standards, involving instruments of information-gathering and behaviour modification" [Black, 2001].*

The holders of this authority are *regulators*. Their role is twofold: to create, and to enforce regulations. The financial sector typically distinguishes between these two functions through the use of two separate terms: regulators and supervisors. **Regulators** are in charge of drafting the rules, and **supervisors** of verifying that the rules are applied. In this thesis, we consider the perspective of supervisors¹⁴ in the domains of customer protection in life-insurance and anti-money laundering.

¹⁴ also called "regulatory supervisors"

Regulations are designed to meet specific objectives. The question of whether it is the pursuit of social welfare that animates regulation has been debated for decades in the economic sphere [Levine and Forrence, 1990, Levi-Faur, 2011]. However, scholars generally agree that regulation can be presented as an instrument to promote the general interest, particularly in situations of market failure [Moss et al., 2009]. For example, some regulations aim to protect customers against asymmetries of information, preserve trade secrets or prevent fraud. In this thesis, we examine specific cases of the use of AI in the highly-regulated financial sector [Hadjjemmanuil, 2015]. We focus on two narrow objectives of financial regulation. We explore the case of protecting customers from the knowledge asymmetry that arises between them and an online recommender system of life-insurance. This is the "customer protection objective" of financial regulation, as presented in [Hadjjemmanuil, 2015]. Additionally, we analyse the issue of preventing money laundering using AI systems,

in which the applicable regulation pursues the "reduction of financial crime objective" [Hadjjemmanuil, 2015].

To enforce regulation, supervisors carry out controls of regulated entities, also known as inspections [Hadjjemmanuil, 2015], in which they verify that the rules are being properly applied¹⁵. Inspections are close to the concept of an audit. However, audits are not necessarily on-site nor carried out by regulators. They are usually conducted by other parties external to the entity being monitored [Wright, 2017]. The literature on *algorithmic audits* has grown in recent years. Audits of AI systems in production in regulated industries have adapted historical approaches to auditing from the social sciences [Vecchione et al., 2021, Sandvig et al., 2014, Metaxa et al., 2021, Mökander et al., 2023].

¹⁵The finance industry is known to impose tight regulatory controls on banks and other financial intermediaries [Hadjjemmanuil, 2015].

Definition

Audit, auditability. *In the context of a regulated environment, an algorithmic audit is a governance mechanism in which auditors participate in a field experiment to diagnose the compliance risks associated with AI systems in relation to specific regulations [Sandvig et al., 2014, Metaxa et al., 2021, Mökander et al., 2023]. The auditability of AI systems enables "the assessment of algorithms, data and design processes" [High-Level Expert Group on AI (HLEG), 2019] and permits auditors to conclude on the compliance of AI systems [Toader, 2019, Raji et al., 2020].*

The EU's High Level Expert Group on AI highlighted the key role of auditability for accountability [High-Level Expert Group on AI (HLEG), 2019]. Koshiyama et al. [2021] give four main verticals of algorithm auditing: performance and robustness, bias and discrimination, explainability, and privacy. Some of these verticals they argue, are "closely linked to the principle of prevention of harm [High-Level Expert Group on AI (HLEG), 2019]." Audits aim to verify that systems do not adversely affect human beings.

This regulatory enforcement process contributes to making regulated firms *accountable* for their AI systems. Doshi-Velez and Kortz [2017] define accountability as:

Definition

Accountability. *"The ability to determine whether a decision was made in accordance with procedural and substantive standards and to hold someone responsible if those standards are not met." [Doshi-Velez and Kortz, 2017]*

According to Kroll et al. [2016], the accountability mechanisms that oversee critical decisions, such as loan approvals, immigration procedures or vote counting, are lagging behind technological advances. The authors argue that new technological approaches are needed to verify that AI-based decision-making processes are accountable and compliant with a set of standards.

Additionally, an important element of accountability is the capacity to demonstrate compliance. Felici et al. [2013] state: "*Accountability involves*

[...] demonstrating ethical implementation to internal and external stakeholders". We consider that this demonstration element is provided by the concept of justification.

Justification is another central concept in the enforcement of regulations. During inspections, regulated entities typically need to justify that their current practices are consistent with applicable regulations and their underlying objectives¹⁶. We adopt the following definition of justification provided by Henin and Le Métayer:

Definition

Justification. According to Henin and Le Métayer [2022], a justification, or "justifiability", is an argumentative process that refers to external norms to argue that a decision (or a system) is "good" (or adequate). Justifications are grounded in norms, such as legal requirements [Henin and Le Métayer, 2022, Hildebrandt, 2019].

This definition works in relation to the decisions of an AI system. Henin and Le Métayer [2022] further defines the concept of *legitimacy* in regards to when the AI system as a whole is "good" within some regulation, objectives or system of norms [Suchman, 1995, Henin and Le Métayer, 2022]. Hereinafter, we use the expression *justifiability* to refer to the adequacy of both an AI decision or whole system with respect to applicable legal requirements [Henin and Le Métayer, 2022], for the sake of simplicity.

To date, little work has addressed the role of explainability in the regulatory enforcement process, *i.e.* for accountability, auditing, or justifiability. Doshi-Velez and Kortz [2017] argued that explanations have an important role in enabling accountability of AI developers and users. The practice of providing reasons for decisions has an important legitimacy function in legal culture, promoting trust of decision-making, the rule of law, and acceptance of outcomes [Schauer, 1995]. However, Henin and Le Métayer [2022] highlighted the fundamental differences between justifications and explanations. Contrary to explanations, which are descriptive and contained to the AI system, justifications are normative and extrinsic¹⁷. Hildebrandt [2019] also states that explanations are not sufficient to justify a decision and that a justification may require an explanation, but not systematically. She adds that "*we must not allow the discourse of explainability to stand in the way of the question whether a decision is legally justified, which requires a specific type of legal reasons*" [Hildebrandt, 2019]. Nevertheless, explanations may be necessary to provide tangible information about AI systems' behavior on which to base legal arguments.

AI explanations, justifications, and audits provide pieces of evidence about the trustworthiness of AI systems. However, the point of view of regulators, who are responsible for auditing and requesting justifications, has not been empirically investigated. This thesis addresses this issue by studying how human-centric explainability can support justifications for AI systems during regulatory inspections, taking the perspective of financial supervisors.

¹⁶ The regulators' need for justification is actually something we hypothesize and document in this thesis. To date, very little work has been done to understand the socio-technical reality of inspections.

¹⁷ In Chapter 6, we argue that justifications must also be grounded in intrinsic and accurate information about AI systems implementation, such as explanations.

Additionally, explainability may also have a role to play in certain trust calibration mechanisms that are critical for compliance. We can distinguish several trust relations that influence the compliance of AI systems with some regulation.

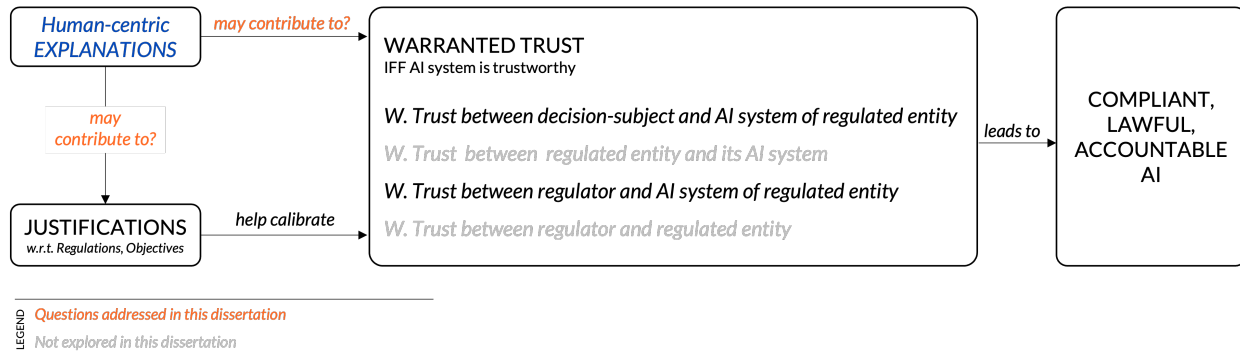


Figure 1.5: The concept of warranted trust and the trust relationships explored in this dissertation. We investigate whether explanations can enhance warranted trust between an individual subject to an AI decision and the AI system, as well as whether explanations can contribute to the development of justifications that support warranted trust between a regulator and the AI system of a regulatee.

First, customer protection regulation may require that customers be able to make informed choices by receiving meaningful explanations about an AI-based recommendation for some product or service. If customers appropriately trust and rely on the recommendations, it indicates that the regulated entity provides users with the necessary means to calibrate their trust in the system, or that the system only provides appropriate recommendations. Either outcome is a sign of compliance.

Second, compliance is often guided by an appropriate level of trust and a healthy dose of skepticism between the regulated entity and its AI system, specifically in high-risk industries. For example, a human AI operator who blindly escalates AI-generated financial crime alerts will be guilty of overtrust, thereby breaching legal requirements about meaningful human review of alerts.

Third, warranted regulator trust in AI systems of regulated entities enables regulator to appropriately assess the legality of AI systems, thereby contributing to "lawful AI". Justifications enable "justified" trust by articulating reasons to trust or distrust an AI system. Explanations and audits are likely to play an important role in supporting such justifications with factual evidence about an AI system's behaviour.

Fourth, regulators' trust in regulatees also influences compliance in a complex and contradictory way [Six, 2013]. On the one hand, if regulators fully trusted regulatees, there would be no need for inspections, and public trust in regulators would be reduced [Six, 2013]. On the other hand, some research has shown that if regulators act out of distrust in regulated entities, the overall result is poorer compliance [Gunningham and Sinclair, 2009]. It has also been shown that the more inspectors trust regulated entities, the more likely they are to be compliant [Braithwaite and Makkai, 1994].

In this thesis, we investigate the first and third trust relationships through two case studies. We examine the challenges to warranted trust

between customers and AI systems in life-insurance, and trust between regulators and regulatees' AI systems in anti-money laundering.

1.1.6 Research domains

Explainability is an interdisciplinary topic. XAI researchers have primarily focused on developing statistical tools to gain insight into the inner workings of "black boxes". For example, many techniques rely on querying the AI system and looking at specific entry/outcome pairs. Varying degrees—local or global—and types of explanations¹⁸ can be achieved. The technique for generating explanations is a critical research stream where much progress has yet to come on the robustness, fidelity, causality of explanations. However, other fields of research like human-computer interaction (HCI), social sciences or law help us make sure we keep this research aligned with why we want to generate explanations and what kind of explanation is needed in specific situations, *i.e.* the human and societal aspects of XAI [Longo et al., 2020]. For instance, a new stream of research called "contestable AI" [Alfrink et al., 2023, Balayn et al., 2023, Lyons et al., 2021, Kaminski and Urban, 2021] aims to design explanations for citizens to contest an algorithmic decision. In recent years, an increasing body of research has been dedicated to studying people' needs for explanations, relying on qualitative user studies or on cognitive science theories. It has also endeavoured to better understand the effects of explanations on users to inform their design.

¹⁸ For example, counterfactual explanations explain the minimal changes to make for a specific decision to be flipped.

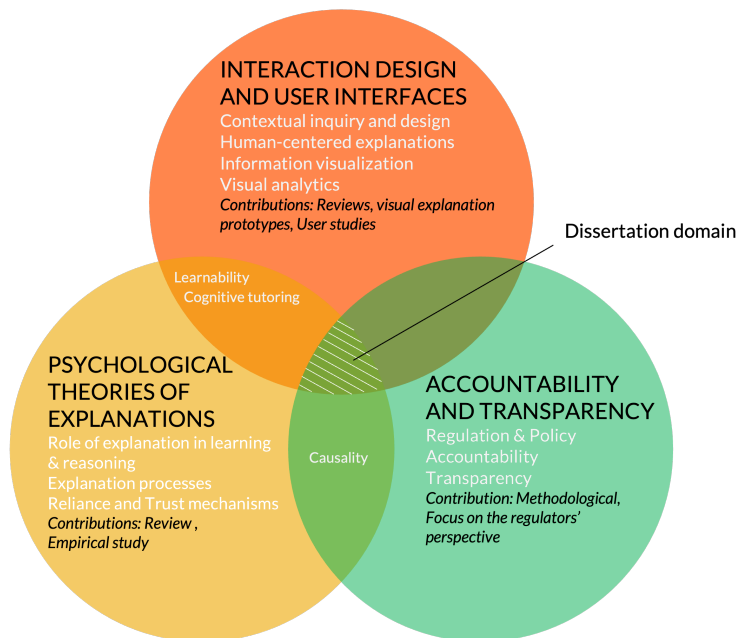


Figure 1.6: Domain scope

The work presented in this dissertation falls within this line of research. It is situated at the intersection of three primary research communities, all focused on the subject of explainability: (1) the design of interactive interfaces, rooted in HCI, (2) psychological theories of explanations, (3) and the study of algorithmic fairness, accountability and

1.2 Problem statement

Explanations of AI systems are supposed to lift the veil of AI's complexity, enable meaningful human understanding, and solve the black-box problem. However, their effects on warranted trust, and specifically the warranted trust of regulators and customers to enable compliance, has not been clearly established [Poursabzi-Sangdeh et al., 2019, Wang et al., 2019a, Ghassemi et al., 2021, Kaur et al., 2020]. For example, Ghassemi et al. [2021] argue that explainability is a "false hope" in health-care to help inform patients and Kaur et al. [2020] showed that the data scientists in their experiments relied too heavily on XAI tools overall, and used them to rationalize suspicious observations. Additionally, it seems clear that explanations are bound to fail if they are not "human-centric", *i.e.* tailored to their human audience and purpose [Tomsett et al., 2018, Ooge, 2023, Ooge et al., 2022, Maxwell, 2023]. Various groups, such as medical doctors or AI practitioners (AI developers or expert users), have received a certain amount of attention in the literature [Wang et al., 2019a, Ghassemi et al., 2021, Panigutti et al., 2023a, Sun et al., 2022, Liao et al., 2023]. However, there is a scarcity of research on the development of human-centric explainability addressed to either customers or regulators to verify compliance. The importance of this issue is likely to increase in the future as more regulations are introduced.

The question we address in this dissertation is: *To what extent can human-centric explainable AI enable warranted trust and regulatory compliance in financial applications?* To answer this question, we break down the problem into two parts:

PROBLEM 1 *How do AI explanations affect our trust calibration in AI predictions and systems?* As we presented in Section 1.1.4, it is still unclear whether AI explanations are able to lead to warranted trust. Specifically, some argue that explanations can lead to various cognitive pitfalls, leading to inappropriate trust and poor decision-making [Chromik and Butz, 2021, Ghassemi et al., 2021, Kaur et al., 2020]. This research therefore begins with the identification of what cognitive patterns might get in the way of appropriately using, interpreting and trusting explainable AI decision systems. We first ask: *What are the cognitive challenges to fostering appropriate trust in explainable AI?* We review the cognitive biases that intervene in the trust calibration process, notably those that lead to overtrust or distrust of AI decisions. We stress the need for human-centric XAI design, that take into account human cognitive constraints. Secondly, in response to a growing interest for designing more interactive explanations [Weld and Bansal, 2018, Cheng et al., 2019], we examine whether interactive explanations designed to fit the human cognitive architecture are more effective in enabling warranted trust. We ask: *To what extent can "human-like" interactive explanations help overcome trust calibration issues?*

PROBLEM 2 *To what extent can explainability support regulatory compliance of AI in the financial sector?* Although AI is increasingly entering regulated industries and new AI regulation is emerging [European Commission, 2021], very little research has examined the role of explainability to ensure regulatory compliance. In the second part of this dissertation, we examine how AI explainability can foster warranted trust by customers, and warranted trust by regulators, and thereby meet regulatory objectives in two applications of AI in finance. In the first case study, customers' warranted trust in an online recommender system of life-insurance contracts is a desirable objective of customer protection regulation. We therefore ask: *Does explainability enhance customer warranted trust and empowerment in life-insurance?* We also ask: *What is the impact of different explanation formats, including interactive ones, to meet this regulatory objective?* In the second case study, we examine the role of explanations to enable the warranted trust by regulators to evaluate compliance of AI systems in anti-money laundering and countering terrorism (AML-CFT). Our research question is as follows: *What are the regulatory supervisors' needs for explainability to justify the decisions and characteristics of AI systems in AML-CFT?*

1.3 Thesis overview

This dissertation is divided into seven chapters, including this introduction, and two research parts. *Chapter 2: Background* reviews the relevant literature setting the stage for this research. In particular, it sheds light on the different disciplinary approaches in the very active field of explainability, which has grown impressively in recent years. To examine the challenges of human-centric explainability in supporting warranted trust and compliance, we then divide our analysis into two parts.

Part I: Calibrating trust in explainable AI: common pitfalls and the promise of interactivity focuses on the cognitive challenges for warranted trust in human-centric explainable AI, taking a cognitive approach. As the field of explainability has grown considerably in recent years, with thousands of academic papers published each year, reviews are much needed to distill important insights. This is why we decided to begin this research with two reviews of the literature. Part I therefore contains two chapters presenting two reviews.

Chapter 3: Trust, overtrust, distrust in explainable AI: a cognitive approach identifies the cognitive processes people use when calibrating trust in XAI-assisted settings, highlighting common uses, misuses and disuses of explanations. We also review the other ways in which cognitive biases affect the design and evaluation of explainable AI.

Chapter 4: Towards "human-like" explanations: the promise of interactivity explores the potential of interactive XAI to limit biases by adopting a more "human-like" explanation process. We present a taxonomy of the different ways in which explanations are interactive and summarise the effects of explanations on trust, reliability or understanding.

Part II: Complying with regulation using human-centric explainable AI: two case studies in finance explores two real-world contexts in finance where explanations may be necessary for compliance. Part II also contains two chapters. The two case studies also provide information on the entry of cross-sector AI regulation, such as the forthcoming AI Act, into a highly regulated sector. The first AI application in life-insurance distribution is considered as high-risk under the AI Act. It is still uncertain whether the second case in AML-CFT is considered high-risk under the AI Act, as the final text of the AI Act has not yet been released, but it is probable. In either case, the study documents how regulators are adapting to AI in light of existing financial regulations.

Chapter 5: Empowering customers of robo-advisors with explainability investigates the explanation needs of customers of life-insurance robo-advisors²⁰, and the explanation requirements from the perspective of customer protection supervisors in this context. We examine, in a controlled study, the correspondence between the regulatory objectives of explanations and their actual effects on users. Specifically, we focus on explanations' effect on appropriate trust and reliance by customers. We test different forms of explanations, including interactive ones. We highlight the challenges that arise to empower users while avoiding misplaced trust.

²⁰ A robo-advisor is an online platform for financial investment advice.

Chapter 6: Understanding the supervisors' needs for explainable AI in financial crime detection analyses the needs of regulatory supervisors for explanations to audit AI decisions and systems using a qualitative workshop-based method and a legal analysis. This user-centric approach allows us to delineate the challenges of using explainability for demonstrating compliance in AML-CFT. We also describe the socio-techno-legal context of supervisors and their auditing practices in this domain.

Chapter 7 concludes on the main findings of this thesis and discusses open questions and avenues for future research.

1.4 Research approach

Human-computer interaction researchers are concerned with observing how people interact with tools that they build. Explainability research also involves designing XAI artefacts and observing users interacting with them in context. This dissertation applies a set of behavioural research methods to collect information on user behaviour with XAI. The studies conducted in this work follow typical methods used in explainability and HCI research, such as reviews, and field experiments. We also demonstrate the usefulness of bridging legal and HCI approaches. Our argument is that a comprehensive understanding of the legal requirements enforced by regulators is necessary to understand the needs of this user group. Below is a brief description of the methodological approaches we employed for observing and designing (in *italics*) human-XAI interactions.

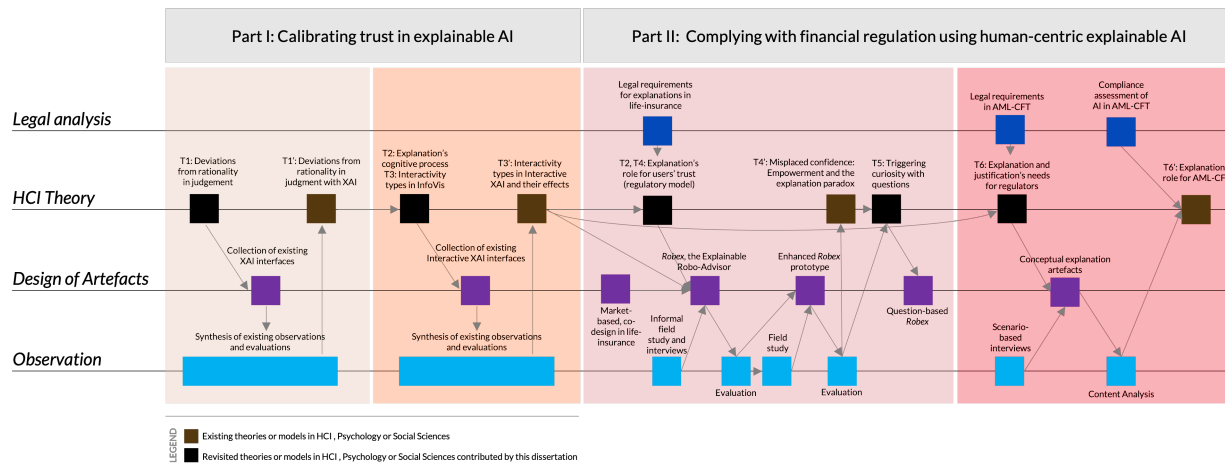


Figure 1.8: Overview of the work presented in this dissertation through a modified version of the triangulation framework of Mackay and Fayard [1997], inspired from [Huron, 2014]

Mackay and Fayard [1997] described a triangulation framework which explains how natural sciences, design and engineering sciences can be integrated. We present an adapted triangulation framework in Figure 1.8 showing the contributions of our work.

Reviews,
Collections

At the start of my PhD, a significant number of primary studies on explainability had been freshly published but there was little hindsight or analysis about them. Thus, it seemed fitting to synthesize that work through literature reviews. We used detailed scoping reviews in Chapter 3 and 4 to synthesize some observations made on XAI-human interaction. Scoping reviews are an appropriate survey type to examine how research is conducted on a specific topic, give a summary of the focus of the field, map key concepts, identify the types of evidence found in a field, pave the way for future systematic reviews, and identify gaps in the literature [Munn et al., 2018]. Moreover, reviews are also ways to get inspiration for the design of XAI artefacts [Herring et al., 2009]. This *collection* process allows to identify state-of-the-art designs as well as features that do not exist yet. In both of the reviews presented in this paper, we followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyse) methodology, a systematic, standardised way of collecting papers [Page et al., 2021, Tricco et al., 2018].

Co-design

In the context of life-insurance (Chapter 5), we first relied on a market-driven approach to understand the complexities of the domain, and take inspiration from existing online life insurance tools to make our experiment as realistic as possible. We then supplemented our approach with interviews with regulators and non-expert users to enhance our understanding of the life insurance industry and end-users needs, following a co-design methodology [Panigutti et al., 2023a, Luria, 2023]. Co-design in the context of human-computer interaction (HCI) refers to a collaborative and participatory approach [Spinuzzi, 2005], where both researchers and end-users engage in the design process. This approach recognizes the importance of involving users to meet their needs, preferences, and expectations effectively.

Field experiment

This method involves investigating the impacts of a phenomenon with some controlled variables, but in a real-world setting. Mcgrath [1995] describes it as *“working within an on-going natural system as unobtrusively as possible, except for intruding on that system by manipulating one major feature of that system.”* It offers the advantage of increased generalisability, enabling testing with a larger number of participants, while minimising invasiveness. Nevertheless, it sacrifices a certain degree of control.

In this Chapter 5, we wanted to study the effects of different formats of explanations on regulatory objectives, including user understanding. As the case study dealt with robo-advisors, which are online platforms, we decided to conduct an online field experiment using a crowdsourcing platform to recruit potential users. This approach curtails the invasive impact of the research.

Interviews

In this work, we developed interview protocols several times to better understand our case studies contexts and stakeholders' needs. Interview guides can be found in the Appendix. Our first set of interviews were conducted in life-insurance (Chapter 5), where we used semi-structured interviews with a think aloud section in which participants used our explanation prototype. We chose this approach to better understand and compare the perspectives of different user groups and improve our explanation prototypes. In Chapter 6, we conducted interviews again, in the context of anti-money laundering. Our aim was to gain an in-depth understanding of the regulators' perspective. As a result, we opted for focus groups [Morgan, 1996], using a semi-structured interview protocol based on scenarios. Each time, we took a grounded analysis approach, as described in [Creswell, 2012], either using simple thematic coding or by combining it with axial coding.

Compliance assessment

In our AML-CFT case study (Chapter 6), we observed that the interview participants, particularly the supervisors, consistently referred to legal requirements or regulatory sanction cases when asked about the questions they had about the AI systems and the explanations or justifications they wished to see. This prompted us to find out more about the AML-CFT laws that participants referenced. We also found that the literature was not clear about how compliance in this domain could be affected by AI's opacity. We therefore supplemented our HCI, qualitative, interview-based approach with a qualitative compliance assessment, *i.e.* a legal analysis. We begun with a doctrinal research as described by McConville [2017]. We highlight in this work the benefits of combining these HCI and legal qualitative research approaches.

1.5 *Major findings*

This section serves as an executive summary of the contributions of this thesis, which are developed in Chapter 7 concluding the dissertation.

1. Explanations tend to increase trust, including overtrust, depending mainly on users' knowledge and skills, and explanations' completeness, framing and timing.
2. Interactive explanations of AI systems tend to increase trust, but not necessarily overtrust.
3. Interactive explanations seem to be more useful for performing a task than static ones, but they are less easy to use and take longer.
4. In the context of life insurance robo-advisors, explanations—even interactive ones—were of little use in helping customers understand algorithmic recommendations and trust them appropriately, thus failing to meet their main regulatory objective.
5. Dialogic explanations provided in natural language (in the form of a chat) increased unwarranted trust of customers in algorithmic recommendations, in the context of life insurance.
6. In the context of anti-money laundering, regulatory supervisors require *justifications* in order to verify: (1) human alignment with AI systems parametrization, (2) business expert understanding of the outputs, and (3) control of AI-specific risks.
7. Explanations have a role of "trial evidence" for justifications. Justifications should not only be extrinsic by referring to norms or regulations [Henin and Le Métayer, 2022], but also intrinsic by depending on faithful evidence of the system's behavior, that explanations can provide.

1.6 Academic publications

Below is an overview of the publications in workshops, conferences and journals that I have contributed to during my PhD.

Publications as first author

"How Cognitive Biases Affect XAI-Assisted Decision-Making: A Systematic Review", Astrid Bertrand, Rafik Belloum, James R. Eagan, Winston Maxwell, Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22), Oxford, UK, August 2022 <https://doi.org/10.1145/3514094.3534164>

"On Selective, Mutable and Dialogic XAI: A Review of What Users Say about Different Types of Interactive Explanations", Astrid Bertrand, Tiphaine Viard, Rafik Belloum, James R. Eagan, Winston Maxwell, Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23.), Hambourg, Germany, April 2023 <https://doi.org/10.1145/3544548.3581314>. Honorable mention.

"Towards Informed Decision-making: Triggering Curiosity in Explanations to Non-expert Users", Astrid Bertrand, 2022 Workshop on XAI and HCI, IHM Conference, Namur, Belgium, April 2022 <https://hal.science/hal-03651368/document>.

"Questioning the ability of feature-based explanations to empower non-experts in robo-advised financial decision-making", Astrid Bertrand, James R. Eagan, Winston Maxwell, Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), Chicago, USA, June 2023 <https://doi.org/10.1145/3593013.3594053>.

To appear: "AI is Entering Regulated Territory: Understanding the Supervisors' Perspective on Model Justifiability in Financial Crime Detection", Astrid Bertrand, James R. Eagan, Winston Maxwell, Joshua Brand, conditionally accepted for publication in the proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24), Honolulu, Hawaii, May 2024.

Publications as co-author

"Do AI-based anti-money laundering (AML) systems violate European fundamental rights?", Winston Maxwell, Astrid Bertrand, Xavier Vamparys, International Data Privacy Law, Volume 11, Issue 3, August 2021, Pages 276–293, April 2021 <https://doi.org/10.1093/idpl/ipab010>.

"Are AI-based Anti-Money Laundering (AML) Systems Compatible with European Fundamental Rights?", Winston Maxwell, Astrid Bertrand, Xavier Vamparys, ICML 2020 Law and Machine Learning Workshop, Vienna, Australia, July 2020 <https://hal.science/hal-02884824/document>.

Chapter 2

Background

THIS CHAPTER provides an overview of the explainability field focusing on its origins, its interdisciplinarity, and its ongoing and future directions. In Section 2.1, we look at a historical perspective of explainability to reveal the interdisciplinary and far-reaching roots of this emerging field of research. Further, Sections 2.2, 2.3 and 2.5 develop the ongoing work on explainability respectively from a computer science, social sciences and legal angle. Finally, we explore the role of Human-Computer Interaction, as a multidisciplinary field by essence, to advance research in explainability in Section 2.4.

2.1 *A historical perspective on explainability*

Explainability is not a new subject. Before the research interest in explainability erupted in the context of deep neural networks, a wide range of work already existed on the epistemology of explanations and early computational systems. However, Atakishiyev et al. [2020] and Longo et al. [2020] noted the lack of a confirmed and resilient connection between the historical origins of XAI and present-day AI applications. Nevertheless, we can broadly trace back the origins of XAI to two historical avenues: on the one hand, the philosophical and social foundations of explanations; on the other hand, the development of expert systems and machine learning applications¹.

The first historical root of XAI is work on formal theories of explanations. This line of thought challenges us to think about what counts as an explanation, particularly in science, and what purposes explanations serve. Throughout the evolution of philosophical thought, scholars have analysed the nature and types of explanations, their explanatory power, functions, and reach [Bunge, 1998, Lombrozo, 2006, 2016, Hilton, 1988]. Aristotle already discussed the notion of explanation [Falcon, 2006], arguing that "*knowledge becomes scientific when it tries to find the causes of why*" [Longo et al., 2020]. This has been reiterated in more recent literature, which emphasises the challenge of responding to "why-questions" that entail counterfactual and abductive reasoning [Pople, 1973, Muggleton, 1991, Poole et al., 1987, Miller, 2019]. Counterfactual reasoning involves testing whether an event E is the cause of a phenomenon of interest P

¹Expert systems are usually regarded as the first implementation of AI [Russell and Norvig, 2010].

by mentally undoing E and assessing how it affects P [Miller, 2019]. Abductive reasoning originates from the field of formal philosophy and involves constructing an explanation that best fits a set of observed data [Atakishiyev et al., 2020, Longo et al., 2020, Miller, 2019]. It is often described as "inference to the best explanation" [Harman, 1965]. This strand of work has also stressed the importance of causality in explanation [Halpern and Pearl, 2005]. For example, Hilton [1988] established the notion of "causal chain", *i.e.* successive causes that lead to the occurrence of the phenomena of interest. Meanwhile, other work in social sciences highlighted the structural and social aspects of explanation [Roth, 1989, Malle, 2004, Graaf and Malle, 2017, Miller, 2019].

As Longo et al. [2020] highlighted, little connection has been made so far to the formal history of XAI, *i.e.* theories of explanation or causation [Holzinger et al., 2019]. Miller [2019]'s review stands out as a rare work that links this knowledge in philosophy and social sciences to modern applications of AI. This introductory paragraph on the study of explanation in the fields of philosophy, sociology and psychology only scratches the surface of the vast body of knowledge that has accumulated on the subject over the centuries. We will develop the important findings from these disciplines in Section 2.3.

The origins of explainability as a field of research can also be linked to an early body of work on the explanation of socio-technical systems dating back as far as the 1950s. As soon as computers became more sophisticated and "intelligent", thanks to the implementation of knowledge and rule-based reasoning in expert systems, the question arose of how to explain their decision-making procedures in a synthetic and comprehensible way that is adapted to the explanation recipient. From this point of view, explainability is nothing new. For example, the book by Winograd and Flores published in 1987 "Understanding Computers and Cognition" examines the underpinnings of understanding what computers do, in relation to human language, thought, and action. A seminal, early work on the design of explanations for expert systems can be found in medicine [Confalonieri et al., 2021]. MYCIN was a famous expert system designed to assist doctors in their diagnosis about infections. It was presented by Buchanan and Shortliffe in the 1970s. The system was based on domain and factual knowledge modeled as "production rules". It was able to provide explanations as "lines of reasoning" of the system [Confalonieri et al., 2021], that is to allow the user to explore the sequence of rules that were used. Moreover, it included a question-answering module, allowing the user to seek answers for some predefined questions.

Other expert systems, featured explanation as "stories", presenting how a system considered a problem and some observations, then inferred hypotheses, studied causal relations and eventually found a cause for the problem [Confalonieri et al., 2021, Roth, 1989]. An example of expert system presenting such reasoning is Rex [Wick and Thompson, 1992]. It used a story structure, a set of reasoning cues, problem and solutions constraints to produce explanations.

y-axis represent the number of citations of the article, and its colour indicates its discipline. This graph was made by searching for many different keywords related to explainable AI and interpretable AI³ on the Semantic scholar database, which enables to sort results per citation count. The top 60 was refined by plotting the citation graph for a few papers in Connected Papers. We stopped collecting papers when we did not find any new addition to the top 60 when we plotted different graphs or searched for different XAI-related keywords.

We can see that the green bubbles, representing the computer science field, are far more numerous and wider in this top 60. Popular papers—with over 10000 citations—looked at interpretation of convolutional neural networks (image classifiers), like [Simonyan et al., 2014], which presented gradient-based saliency maps, or [Zeiler and Fergus, 2013] which introduced a feature visualization in ConvNets. Other seminal work, like [Lundberg and Lee, 2017] and [Ribeiro et al., 2016], presented techniques to identify the most important features used by any kind of classifier. Meanwhile, [Doshi-Velez and Kim, 2017], [Adebayo et al., 2020] and [Kim et al., 2016] provided critical introspection into the emerging field of XAI, but still focused on the computer science side.

Contributions in legal and social sciences are scarce in XAI, comparatively to computer science, as shown in Figure 2.2. However, interdisciplinary work is gaining traction. For example, [Lipton, 2018], [Burrell, 2016] or [Rudin, 2019], reflect on the discourse of interpretability, on the problem of opacity, or on the use of inherently opaque models vs. interpretable ones. For example, Lipton [2018] highlights that *"Papers provide diverse and sometimes non-overlapping motivations for interpretability, and offer myriad notions of what attributes render models interpretable"*. Kulesza et al. was a pioneer in studying explainability from an HCI lens [Kulesza et al., 2013, 2015]. Starting from approximately 2018-2019, XAI gained popularity among HCI researchers. They have focused on better understanding users' needs, designing user-centered XAI interfaces or developing user-centered metrics for evaluating XAI [Wang et al., 2019a, Hoffman et al., 2019].

Given the exponential body of work in explainability, review papers have been timely contributions in recent years to process important insights and to navigate the myriad of XAI techniques, XAI design artefacts, evaluation metrics, or XAI goals and applications. Seminal review work include [Adadi and Berrada, 2018, Barredo Arrieta et al., 2020, Abdul et al., 2018, Carvalho et al., 2019, Miller, 2019, Guidotti et al., 2018]. This dissertation contributes to this need for review papers in the Chapters 3 and 4.

³ For example *"interpretable AI system", explainable machine learning, explanation algorithm, trustworthy AI, etc.*

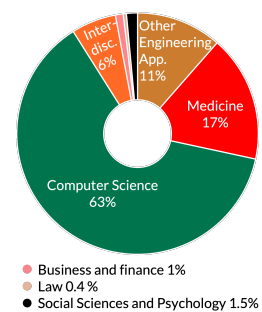


Figure 2.2: Distribution of contributions in explainable AI across disciplines. This graph is based on a corpus of 5756 articles published from 2015 to present, extracted from searching "explainab*" in the article title in the Scopus Database.

2.2 Explainability in Computer Science: the toolbox

Explainability for modern AI applications has first been approached as a purely technical problem. The aim was to find tools to meet computer scientists’ growing interest in understanding what happens in neural networks [Atakishiyev et al., 2020]. As a result, a myriad of explainability techniques have been proposed over the last ten years. We provide a brief overview of these in Section 2.2.1. In Section 2.2.2, we review the current technical challenges for generating and evaluating explanations in machine learning.

2.2.1 The wide range of explainability methods

The array of explanation techniques provided by the computer science community is extensive. This breadth arises from the wide scope of machine learning which encompasses diverse data types, such as images, text, tables, audio, graphs, and time series, as well as a range of models, spanning from DNNs, Bayesian Networks, SVMs, to Tree Ensembles. Moreover, there are varying approaches to the explainability problem. For instance, explanations may pertain to *specific* data and model types or be *agnostic* and applicable to any model or data. Another possibility is for explanations to be *local*, focusing on individual forecasts, or *global*, offering a comprehensive explanation of the model throughout its definition range. Further, explanations may arrive *post-hoc*, meaning that an explanation is reconstructed given some inputs and predictions from a model. This is also known as reverse engineering in the literature [Guidotti et al., 2018]. However, explanations can also be *built-in*, meaning that the model is trained in a way that is inherently interpretable (*e.g.* white box models, training with sparsity constraints or with supervised explanations). Many surveys have proposed taxonomies to gain a clearer picture of the different types and approaches of explanations [Barredo Arrieta et al., 2020, Guidotti et al., 2018, Nauta et al., 2023, Burkart and Huber, 2021, Carvalho et al., 2019, Das and Rad, 2020, Mohseni et al., 2021b, Molnar, 2019, Gilpin et al., 2018]. We drew on these to summarize main explanation concepts, production mechanisms and representations, using our synthetic categorization outlined in Figure 2.3.

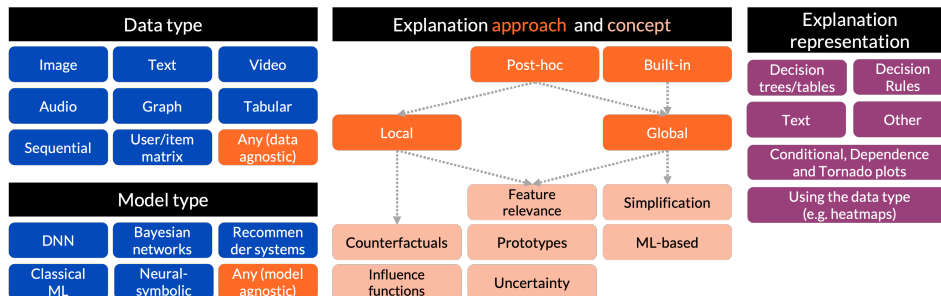


Figure 2.3: Categorization of explainable AI methods along four dimensions inspired by Nauta et al. [2023] and Barredo Arrieta et al. [2020].

Explainability methods and production mechanisms

Let us begin with an overview of *what* are the explanations offered by the computer science field. We present below six different explanatory concepts that can help inform on the operation and behaviour of black-box models.

- **Feature-based.** One of the most popular way to shed light on AI models' inner workings is by determining the influence of input features on the outcomes or intermediate representations of the model. We include in this category feature importance, feature attribution, activation maximization [Nguyen et al., 2016] and saliency methods [Zeiler and Fergus, 2013].

Feature importance consists in generating a vector with the weight and magnitude of the inputs used by the black-box. It can be either local or global. It is also sometimes referred to as *feature attribution* such as in [Lundberg et al., 2019] for tree ensembles. There are various approaches to creating this vector, such as using game-theory inspired computations [Lundberg and Lee, 2017] or using the coefficients of a linear model that approximates the black-box in a region of interest [Ribeiro et al., 2016]. Most of these methodologies are post-hoc and rely on querying the black box using input records produced in a controlled manner or through random perturbations of the original training or testing data [Guidotti et al., 2018].

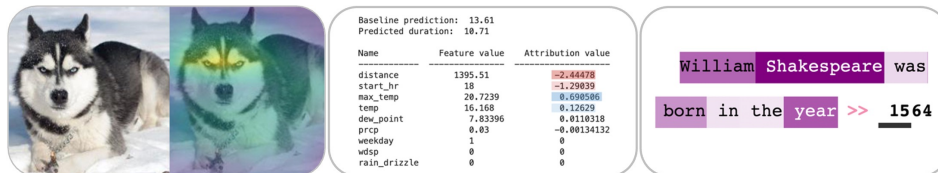


Figure 2.4: Illustrative examples of feature-based explanations for different data types (image, tabular and text data) with input saliency [Alammar, 2021, Unruh and Robinson, 2020].

Saliency methods consist in determining the inputs (either words in a sentence or areas in an image) that are most “salient” from a model’s perspective. They are broadly divided into three categories [Kindermans et al., 2017]. *Sensitivity methods* show how a small change to the input affects the prediction [Simonyan et al., 2014]. *Signal methods*, like DeConvNet [Zeiler and Fergus, 2013] or Guided BackProp [Springenberg et al., 2015], look at the neuron activations [Carter et al., 2019] in the model to attribute importance to input features. This type of method is also know as *activation maximization*. Finally, *attribution methods*, like Integrated Gradients [Sundararajan et al., 2017] aim at completely specifying the attributions for all the input features so that they sum up to the output. Saliency techniques usually rely on gradient-based calculations. To gain a better understanding of gradients, let’s consider a CNN that classifies cats and dog images. Figure 2.5 illustrates how changing individual pixels affects the model’s identification of the picture as a “cat”: the upward arrows represent changes that make it more likely for the model to identify the image as a cat. Additionally, the thickness of the arrow indicates the amount of gradient

shift that occurs due to that pixel being altered. Pixels that alter the image substantially are called "salient" and are usually represented in white or warm colors in saliency maps.

- **Prototype-based** methods consist in extracting representative examples or "prototypes" of the black-box outcomes. This approach is inspired by case-based reasoning, which allows users to reason based on retrieved similar input patterns and their outcomes. However, Kim et al. [2016] argued that "examples were not enough" and can lead to over-generalization. They proposed to also to "criticize" the extracted prototypes by extracting "criticism" samples that are not well-explained by the prototypes. These techniques are based on calculating similarities or discrepancies between distributions. Other methods include finding prototypical parts in images pointing to aspects of one class or another [Chen et al., 2019], finding the nearest neighbors of a point of interest in the input data space, or finding prototypical concepts that represent a class [Kim et al., 2018, Ghandeharioun et al., 2022].
- **Counterfactual explanations.** Algorithms can also be explained by considering how an outcome could be changed to another outcome, for example more desirable [Stepin et al., 2021]. The problem of finding a counterfactual explanation in ML is usually described as "the smallest change to the feature values that changes the prediction to a predefined output" [Molnar, 2019]. This is achieved by defining a notion of distance between the point of interest and a hypothetical point for which the outcome would be different. Counterfactual explanations have received a growing attention in recent years because of their potential to be *actionable*⁴, their alignment with people's needs for explanations: people usually ask for explanations when the AI outcomes violate their expectations [Kizilcec, 2016]. Other advantages include that they do not require model disclosure or place no constraint on model complexity. Barocas et al. [2020], however, warn against the fact that defining a notion of distance to compute counterfactuals is challenging, and implies somewhat arbitrary choices about the normalization of features. Moreover, counterfactual explanations do not examine the rationality or difficulty of recommended actions and may, for example, suggest that an individual should make less money, or stay longer at his current job [Barocas et al., 2020].

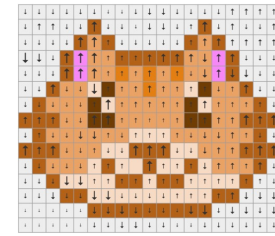


Figure 2.5: Illustration of the gradient-based method to identify "salient" pixels.

More at: <https://pair.withgoogle.com/explorables/saliency/>

⁴Counterfactual explanations help identify features that, if changed can lead to a different result. These explanations are actionable if the identified features can in fact be changed easily. Age or ethnicity for example, are not actionable features that someone can change to get admitted into a school. Getting good grades, however, is an actionable feature.

"One decision maker might scale the axes such that increasing income by \$5,000 annually is equivalent to an additional year on the job. A competing lender, using different training data, could conclude that \$10,000 of income corresponds to one year of work. These lenders might therefore produce different explanations depending on the scaling of attributes."

Extract from [Barocas et al., 2020] on normalizing features for counterfactuals.

- **Influence functions.** Koh and Liang [2020] presented influence functions to link model outcomes to influential training points. Also known as training data attribution, influence methods suggest which training

data points might be the cause of a model's behavior for a given input and output [Pruthi et al., 2020, Akyürek et al., 2022].

- ***Simplification.*** Another popular way to approach explainability is by approximating black-box models by simpler, interpretable ones [Rudin, 2019]. It becomes a problem of "finding an interpretable model that approximates the black-box model as much as possible, typically seeking high fidelity" [Confalonieri et al., 2021]. Those simpler models are called "*surrogate models*". These methods often leverage logical or/and visual models such as decision trees, rules, generative additive models [Caruana et al., 2015], logistic and linear regressions or bayesian models [Kim et al., 2015]. There also exist methods for reformulating "connectionist" models as logical models [Barceló et al., 2020]. This is considered as "built-in" interpretability, which involves setting interpretability constraints like sparsity in the model training [Nauta et al., 2023]. However, the concept of "explainability by design" lacks a fixed set of rules, and the boundaries between an interpretable and a black-box model remain unclear. For instance, it is arguable whether a random forest is typically more explainable than a neural network.
- ***Uncertainty estimation.*** Current explainability methods have often been criticised for their lack of consistency, stability, and for providing little insight into their reliability [Bhatt et al., 2021, Slack et al., 2021, Leavitt and Morcos, 2020, Kindermans et al., 2017]. Consequently, some have proposed to represent the uncertainty of explanations. Slack et al. [2021] proposed Bayesian versions of LIME and KernelSHAP to provide confidence estimates of their quality. Others have considered the uncertainty estimation of black box models as part of the explainability scope [Thuy and Benoit, 2023], or as a necessary complement to transparency [Zhang et al., 2022, Bhatt et al., 2021]. For example, Bhatt et al. [2021] presents different ways, such as Bayesian and frequentist methods, to present uncertainty to stakeholders, that are more accurate than the classic Maximum Class probability method (MCP). Zhang et al. [2022] include both model and explanation uncertainty in their explainability framework.

Explanation representations

Explanations can alternatively be presented in **natural language**, as it was the case for the expert system Rex [Wick and Thompson, 1992], through **plots**, such as partial dependence plots (PDPs), accumulated local effects (ALE) plots, and influence sensitivity plots (ISPs), "tornado plots" that show the feature weights from most to least important, dimensionality reduction plots, through **decision rules, tables or trees** to visually present the logic of the model on specific data ranges, by **leveraging the initial data structure**, such as for saliency maps or prototypes, or by **creating artificial visualizations** of the concept used by, for example, neurons or layers [Nauta et al., 2023]. Above, we have only hinted at the wide range of explanation designs that have been tested. The HCI literature has introduced wide range of explanation visualisations and

interfaces, adapted to the task, context, user, and model at hand. We summarize these efforts in Section 2.4.

2.2.2 *The technical challenges in generating explanations*

Rigorous and falsifiable research. Seminal papers like [Lipton, 2018], [Doshi-Velez and Kim, 2017] or [Leavitt and Morcos, 2020] have warned against a lack of rigor and consensus regarding explainability definition, aims, and practices. In particular, Leavitt and Morcos [2020] noted the growing shortcomings of the methodologies carried out in the XAI literature and endeavoured to analyze them. Specifically, they highlight the lack of scrutiny, criticism and falsifiable hypotheses in explainability research. For example, they point to the incapacity of saliency methods to reflect meaningful properties of the data and network, despite their intuitiveness and appealing visualization [Sundararajan et al., 2017, Adebayo et al., 2020]. More recently, Bilodeau et al. [2023] proved mathematically that some complete and linear feature attribution methods like SHAP or Integrated Gradients do not help more than random guessing for the task of inferring model behavior. The authors point to other, simpler techniques such as repeated model evaluations in order to perform precisely defined interpretability tasks. This critical examination of state-of-the-art research is important for the progress of explainability.

Causality and reasoning. Research in psychology emphasizes the importance of causality in the explanation process [Halpern and Pearl, 2005]. Yet, most of the explanation strategies described above, specifically feature importance explanations, do not provide any measure of causality. Causability is defined in [Holzinger et al., 2019] as *“the extent to which an explanation [...] achieves a specified level of causal understanding with effectiveness, efficiency and satisfaction in a specified context of use.”* It is not because a feature is marked as important that it is necessarily a relevant cause to the outcome. Instead, other confounding variables may be at play. In their review, Confalonieri et al. [2021] noted: *“causal explanations are largely lacking in the machine learning literature, with only few exceptions.”* Consequently, the literature in XAI has been increasingly interested in causal models in search of technical means to address causality in explanations. Explanations based on causal models, like counterfactual explanations, can be *action-guiding*, *i.e.* explain the events resulting from an action [Chattopadhyay et al., 2019, Beckers, 2022]. However, applying causal models to the machine learning field is challenging since it is based on correlation rather than causation [Holzinger et al., 2020, Guo et al., 2021, Peters et al., 2017]. Furthermore, Miller et al. [2017] highlights that identifying causal attributions is not the same as providing a causal explanation, as a complete causal chain is complex and high dimensional, and therefore not comprehensible to a layperson.

Moreover, some have emphasized the shortcomings of XAI to produce explanations based on reasoning and logic. Confalonieri et al. [2021] indicate that *“establishing a common ground of inherent logic from the ground up appears reasonable”*, for example by integrating symbolic or knowledge-based modules in non-symbolic machine learning models. Doran et al.

[2017] also argue that "truly explainable AI should integrate reasoning". Promising work to go in that direction include causal graphs and knowledge graph mining for generating explanations [Holzinger et al., 2021, Lecue, 2020]. These efforts seek either to integrate an external knowledge base (as in [Holzinger et al., 2021]), or to model sets of causes and effects (as in [Lecue, 2020]) in the form of graphs, which facilitates explanation processes.

Evaluation. Additionally, many have highlighted the shortage of controlled and harmonized evaluations of the methods [Longo et al., 2020, Leavitt and Morcos, 2020, Doshi-Velez and Kim, 2017]. This stems from the difficulty of identifying the qualities of an explanation that should be evaluated. The issue at hand pertains to the qualities of a satisfactory explanation, which cannot be resolved by computer science alone. Otherwise, there may arise a risk that AI researchers design explainability for themselves only, rather than for the intended users [Miller et al., 2017]. It has therefore been suggested that evaluations of explanatory agents should incorporate the viewpoint of end-users or a human perspective [Doshi-Velez and Kim, 2017, Miller et al., 2017]. Valuable insights can be gained regarding the quality of explanations through analysis of the social sciences, philosophy, and psychology.

2.3 *Explainability in the Social Sciences: the foundations*

We have already mentioned in this section that the social sciences have played a central, historical role in XAI's pursuit of human understanding. Insights from the social sciences and the philosophy of science [Hedström and Ylikoski, 2010] establish foundational theories regarding the process by which people explain phenomena, or what people look for in explanations. These insights help to bridge the gap between the explainability technique seen in the previous section and the explanations needed to promote human understanding. However, as stated in Section 2.1, contemporary XAI researchers, who have been working on explaining recent forms of AI systems, have been slow to take full advantage of this line of work. Miller et al. [2017]'s review has been a major boost to this endeavour.

Many different aspects of the concept of explanation have been studied in epistemology, philosophy and cognitive sciences, including the reasoning involved in explanations [Lombrozo, 2006, Leake, 1995], the effects of belief and preconditions on explanations [Paul Thagard, 1989], or how people explain the behavior of others, *i.e.* social attribution [Malle, 2004]. All of these facets of explanations are explored in [Miller, 2019].

Below, we make a brief summary of this large array of work, focusing on explanations' role, their contrastive nature, the cognitive and social processes by which we, as human, explain phenomena, and the cognitive biases involved in explanations.

2.3.1 *The role of explanations*

Seeking explanations is part of our everyday life [Williams and Lombrozo, 2010]. Why is my train late this time?⁵ Why didn't you tell your friend? Why is the Earth round? Young children notoriously question literally everything with endless "why?" questions [Williams and Lombrozo, 2010]. In fact, explanations are central to individual's acquisition of knowledge and ability to ascribe mental states to oneself and others⁶[Amsterlaw and Wellman, 2006]. Reasons why people ask for explanations involve assessing the soundness of a claim, support learning, but also satisfy one's curiosity [Miller, 2019]. This study of the role of explanations mainly falls within the domain of philosophy.

Lombrozo [2006] highlights three distinct functions that explanations serve. These functions are 1) to enable the assessment of the likelihood of a claim to be true, referred to as causal inference; 2) to allow for the transfer of knowledge to novel cases, which is known as generalization; and, 3) to assist in the acquisition of knowledge, *i.e.* for learning and discovery. Lombrozo describes how explaining why a claim might be true is an important process for evaluating the soundness of that claim. This process of causal inference often favours mechanistic explanations, *i.e.* explanations of the mechanisms involved in making the explanandum⁷ happen. Further, Lombrozo uncovers how explanation supports generalization. Generalization enables to solve transfer problems and

⁵Very often heard in France. The SNCF, France's leading train company, is often the subject of complaints. That being said French people's reputation for grumbling is accurate.

⁶This ability is known as *theory of mind* in psychology.

⁷The subject of the explanation, or event to explain is called *explanandum* in social sciences.

extend known properties to novel cases. In a controlled experiment, Rehder [2006] showed that participants who were given an explanation for a problem that involved a relevant cause for that problem and another were more able to extend that cause to the other problem than if they did not receive explanations. In short, people could better generalize the cause of one problem to another with relevant explanations. Rehder [2006] also demonstrated that similarity and diversity are important factors in the generalization process. People can generalize from one problem to another specifically when they are similar. Furthermore, people are more likely to generalise an explanation if it stands true in a diverse range of contexts. Finally, Lombrozo argued that explaining novel information to one-self is one of the best ways to learn. Self-explanations are more powerful for learning than *“thinking out loud, reading study materials twice or merely receiving feedback”*. Specifically, self-explaining requires to relate knowledge within prior beliefs [Lombrozo, 2016, Chi et al., 1994].

Similarly, Miller [2019] presented that explanations serve to *find meaning, i.e. to “reconcile the contradictions or inconsistencies between elements of our knowledge structures”*. It has been shown, for example, that people ask questions about events that they find unusual or abnormal [Hilton and Slugoski, 1986]. Additionally, explanations enable us to construct social meaning. Through explanations, we can persuade not only ourselves but also others that a claim is true [Miller, 2019].

Furthermore, we learn more and better when driven by our own curiosity and motivation to understand phenomena. [Shin and Kim, 2019]. Curiosity is driven by an individual’s realization that she has a gap in knowledge, but it decreases if that gap is too large, that is, if the information is unattainable, or if the gap is too narrow, meaning the knowledge is not very useful. In summary, there is an optimal gap in knowledge that maximizes curiosity. Therefore, arousing AI users’ curiosity through explanations is more likely to have an impact.

“The most important factors in the generation of curiosity are an individual’s reference point of knowledge and their awareness of the unknown which is raised by curiosity-evoking stimuli. This information gap then creates a sense of deprivation, which naturally instills a desire to learn.”

Extract from Shin and Kim [2019]

These findings offer valuable insights into people’s needs for explanations. By understanding why we ask explanations and their role in forging knowledge, we can design explanations that people perceive as useful.

2.3.2 *The explanation process*

Understanding *how* individuals explain phenomena to one another is valuable to designing explanations that align with the cognitive architecture of humans. In fact, the *process* of explaining is a defining aspect of an explanation. Miller [2019] describes an explanation as being the *product* resulting from answering a why-question, the *cognitive process* of

inferring plausible hypotheses, probing, selecting and evaluating them, and the *social process* of communicating the explanation to others.

The cognitive process of explaining is composed of several steps, including *causal connection* and *explanation selection* [Miller, 2019]. Causal connection involves identifying plausible causes for an explanandum either through abductive reasoning and/or through simulation. Abductive reasoning involves inferring the most probable causes of an observed event by making hypotheses and testing these. Simulation consists in undoing a likely cause in order to consider the effects of this mutation on the observed event and evaluate the likelihood of the plausible cause actually causing the explanandum. Explanation selection involves selecting a subset of the identified causes, *i.e.* the most "interesting" ones, based on our cognitive biases to discount or regard certain observations. These biases include our attention to causes that are abnormal (unusual causes), intentional (for example deliberate intent is usually seen as a stronger cause for murder than the murder weapon), or functional (causes that cite the function of an object or event). We also tend to select causes that are necessary, sufficient and robust to change.

Explaining is also a social process that follows the conventional structures of a dialogue [Miller, 2019]. Explaining involves the explainee and the explainer asking and answering questions in an iterative way, so that follow-up questions are addressed until the explainee is satisfied. Through this iterative process, conversational explanations are able to be truly relevant by finding the explainee's knowledge gap and taking into account what she already knows. Conversations allow for contextual and incremental explanations [Cawsey, 1993, Miller, 2019].

Social dialogues also have a number of conventions which, if followed, increase the impact and effectiveness of the conversation. These include Grice's maxims of quality, quantity, relation, manner [Grice, 1975]. As Miller [2019] presents it: "Coarsely, these respectively mean: only say what you believe; only say as much as is necessary; only say what is relevant; and say it in a nice way."

Furthermore, according to the theory of mind, which refers to people's ability to attribute mental states to others, individuals engaged in a social explanation process keep track of what has already been explained. Thus, this should also be true for computational XAI agents [Miller, 2019]. In general, the social aspect of explanations calls for XAI agents to also be "socially interactive".

These cognitive and social processes describe the mechanisms employed by an individual (an explainer) *to explain* an event to someone else (an explainee). Explainability leverage these theories to build XAI systems that adopt these processes.

Other useful insights for explainable AI include how people *receive* explanations, as explainees. Miller [2019] details that explanations are evaluated based on our "human" criteria of a "good" explanation. These involve coherence or consistency with prior beliefs [Thagard, 1989, Atakishiyev et al., 2020], simplicity, broadness (or generality) [Lombrozo, 2007], truthfulness and probability. People also prefer explanations that are

simple, *i.e.* which cite fewer causes, and broader explanations, *i.e.* which explain more events [Thagard, 1989, Read and Marcus-Newhall, 1993, Miller, 2019]. Kulesza et al. [2015] highlighted the contradiction of people preferring both simple and complete explanations. However, they found that over-simplification was often problematic for correct understanding of the explained event and suggested to design complete explanations "that do not overwhelm". They also found that completeness was more important than soundness, as it helped participants form more accurate mental models and increased perceived usefulness of explanations.

2.3.3 *Explanations are contrastive*

An important insight from social sciences for XAI put forward by Miller [2019] is that we do not explain an event E , the *explanandum*, per se, but rather explain why E happened instead of some other counterfactual event P . In other terms, in every why-question such as "*why did E happen?*", we ask in reality "*why did E happen, and not F?*" [Miller, 2019, Hilton, 1988]. This is called the *contrastive* nature of explanations. Lipton [1990] refers to E as the *fact*, and F as the *foil*. Miller [2019] presents an illustrative example: if someone in a room asks "Why did Elizabeth open the window?", she surely has a foil in mind that drove her question. There can be many different possibilities for that foil, including "Why did Elizabeth open the door, *rather than leave it closed?*", or "Why did Elizabeth open the door *rather than the window?*". Depending on what the foil actually is, the questions call for different answers .

As [Miller, 2019] or [Stepin et al., 2021] specified, there is a difference between contrastive explanations and counterfactual explanations. Contrastive explanations aim to explain why an output differs from a certain expected result [Miller, 2021], whereas counterfactual explanations point out how to change one result to another. As illustrated in [McGill and Klein, 1993], the former asks "*What made the difference between the employee who failed and the employees who did not fail?*", whereas counterfactual reasoning addresses "*Would the employee have failed if she had not been a woman?*" [Stepin et al., 2021].

2.4 Explainability in HCI: user and context first

While a mathematical perspective is crucial for providing insights into opaque machine learning systems, the social science viewpoint is equally important for offering insights into the human black-box. In turn, the Human-Computer Interaction (HCI) perspective serves as a link between these technical and human aspects of XAI.

In this section, we present the diverse range of contributions that the HCI community provides for the field of explainability.

2.4.1 The need for user-centered explainability

By 1986, Winograd and Flores [1987] had already implemented explanations in early AI systems. They also promoted scientifically-based design principles to replace informal notions of "user-friendly" and "self-explanatory" interfaces. However, these developments have been slow to be transposed to today's AI [Abdul et al., 2018, Broniatowski, 2021]. In recent years, the "modern AI" community has finally begun to recognise the importance of considering the human element of XAI [Longo et al., 2020]. Several computer scientists have advocated for increased human involvement in the process of explanation evaluation [Poursabzi-Sangdeh et al., 2020, Doshi-Velez and Kim, 2017, Vaughan and Wallach, 2020]. These calls were primarily concerned with examining the impact of explanations on users, and evaluating whether specific explanation methods were successful in translating abstract information used by AI systems into human concepts [Doshi-Velez and Kortz, 2017, Kim et al., 2018].

Concerns have also been raised that AI explainability tools are only aimed at computer scientists and are too technical for non-experts and end users to understand, in practical cases of AI development [Miller et al., 2017, Confalonieri et al., 2021, Bhatt et al., 2020]⁸ These discussions encouraged XAI researchers to consider what information end-users actually want and how to present that information depending on the user's context, background, experience and other characteristics.

⁸ As Confalonieri et al. [2021] argues, "aspects of understandability of explanations for lay users has for a long time been overlooked".

The methods, goals and experience of the HCI community in dealing with behavioural research are perfectly suited to this purpose. In fact, the issue of explainability is profoundly a matter of human-computer interaction. By enabling users to make full use of machine learning predictions and systems, explainability aligns with the founding goals of the HCI discipline, which are to expand the range of possible human-computer interactions and collaborations. The HCI community has been studying for decades [Longo et al., 2020] how people interact with computers, how to adapt to users' experience and cognitive architecture, and how to design usable, useful and empowering interfaces [Oulasvirta et al., 2022, Amershi et al., 2019]. Specifically, HCI researchers have drawn heavily on phenomenology and cognitive science to design computer systems and interfaces tailored to the cognitive architecture of the human mind.

The explainability research line adopting an HCI perspective has been

labelled as user-centered or human-centric explainability [Liao and Varshney, 2022]. This approach has made significant progress across various fronts in order to make AI systems more understandable to end users [Wang et al., 2019a, Kim et al., 2018, Liao et al., 2020, Liu et al., 2021, Shin, 2021]. Below we outline 6 main research threads in HCI and XAI research: 1) characterizing explainability user profiles, 2) understanding users' goals and mental states contextually to inform their precise needs for XAI, 3) designing explainable interfaces through iterative cycles of ideation-design-evaluation, 5) developing metrics for evaluating explainability systems and 6) better understanding the factors that contribute to appropriately trust (X)AI systems. The following sections provide an overview of the research advances along these six dimensions.

2.4.2 Different audiences, different goals

The literature in explainability has identified various user profiles [Kirsch, 2017, Rosenfeld and Richardson, 2019, Tomsett et al., 2018, Mohseni et al., 2021b, Langer et al., 2021, Ferreira and Monteiro, 2020]. This helped to identify the gap between the technical explanations provided by the computer science community and the diverse explanation needs of other, real-world XAI users.

Some studies have placed emphasis on AI expertise and application domain knowledge to classify users. As a result, three distinct user groups have been put forward in the XAI literature: **AI novices**, also known as non-experts or lay users, **domain experts**, and **AI experts** [Mohseni et al., 2021b, Ribera and Lapedriza, 2019]. AI novices are individuals impacted by AI systems, but who have little knowledge in the technicalities of AI. Examples are users of an online recommender system, decision-subjects of a loan application, medical patients or citizens interested in learning more about public AI systems. Domain experts are people with significant knowledge in the field of application of the AI system, such as doctors, or loan officers [Ooge, 2023]. AI experts are machine learning developers, engineers and researchers. This classification, however, is coarse. For example, the lay user group is extremely diverse specifically in terms of familiarity with AI [Liao and Varshney, 2022].

Other classifications distinguish user groups according to their role in the machine learning ecosystem. For example, Tomsett et al. [2018] identified six roles that require different, if any, AI explanations, as shown in Figure 2.6. Similarly Hind [2019] identified four explainability user groups: AI system builders, who want to debug their models and test them before deployment; end-user decision makers, who use the AI recommendations to make a decision; regulatory bodies, in charge of protecting citizens' rights; and end consumers, who are directly impacted by the decision of the AI system and may want to contest it. Maxwell [2023] recognises roughly the same four audiences: **machine learning engineer**, **human operator of the system**, **person affected by the algorithmic decision** and **judge, auditor or regulator**.

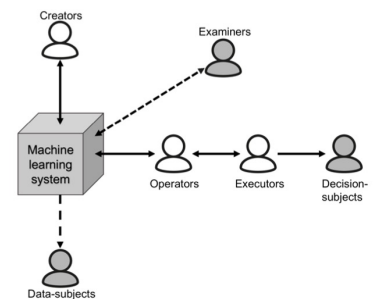


Figure 2.6: Figure 1 in [Tomsett et al., 2018] identifies the different stakeholders in a machine learning ecosystem. "Direction of arrow indicates direction of interaction."

These user profiles ("**whom to explain?**") are associated with different goals ("**why explain?**") [Leake, 1991], leading to different information needs and different explanation designs. Some have listed the explanation content ("**what to explain?**") and explanations methods ("*how to explain?*") corresponding to different user groups [Liao and Varshney, 2022, Liao et al., 2020, Ribera and Lapedriza, 2019, Mohseni et al., 2021b]. For example, Liao et al. [2020] provided a "question bank" of user questions related to explainability, including questions related to the general model logic, "how?", to the changes that would get the alternative prediction, "why not?", or to the feature(s) that if changed, could alter the prediction in a direction, "how to be that?".

The amount of time each user is prepared to invest in the explanation ("**how long to explain?**") [Gajos and Mamykina, 2022, Stumpf et al., 2009] also depends on the user's profile, as does the time at which the explanation is presented to the user ("**when to explain?**"). Nourani et al. [2021] argue that the timing of the presentation of explanations, either before, during or after the explainee has generated her own explanation, greatly affects the user's mental model and reliance on the AI. Maxwell [2023] depict four different contexts in which user attend explanations. These are testing the system, human-in-the-loop, human-on-the-loop or ex-post investigation. Some work also posits different levels of explainability for specific audiences and contexts taking into account legal, economic, social and technical considerations [Beaudouin et al., 2020, Dupont et al., 2020, Langer et al., 2021].

Adadi and Berrada [2018] identified four main reasons *why* people need explainability: **explain to justify** that an AI decision is good, for example to regulators; **explain to control** and identify errors quickly, for example to human operators of the AI system; **explain to improve** AI models, which is what AI developers want; and **explain to discover** new knowledge from powerful AI systems, such as how AlphaGo beats humans at chess. Suresh et al. [2021] and Mohseni et al. [2021b] present **explain to build trust** as a distinct important user goal, specifically for novice users. Suresh et al. [2021] also identified compliance with regulations as a key objective of explainability users, that is tied to the overarching goals of building trust and understanding AI models. More fine-grained and contextual approaches are needed, however, to understand the precise needs of users.

2.4.3 Understanding user needs in context

A growing number of XAI systems have been developed for specific users in specific contexts, with some examples provided in [Zhu et al., 2018, Wang et al., 2019a, Panigutti et al., 2023a, Krause et al., 2016, Coppeters et al., 2018, Cheng et al., 2019, Ooge, 2023]. To provide relevant explainability designs, these studies go through the endeavour of understanding the specific needs of users to support them in their context-specific tasks and goals. HCI researchers have relied on cognitive theories about how users explain [Miller, 2019, Wang et al., 2019a, Bertrand et al.,

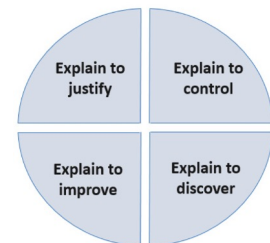


Figure 2.7: The four reasons motivating the need for explainable AI presented in [Adadi and Berrada, 2018].

Who?	<i>Developers or AI researchers / Domain experts / Lay users</i> [Ribera and Lapedriza, 2019, Mohseni et al., 2021b] <i>AI Creator / Operator / Executor / Data-subjects / Decision-subject / Auditors</i> [Tomsett et al., 2018] <i>Machine Learning Engineer / Human operator / Person affected / Regulator or auditor</i> [Maxwell, 2023, Hind, 2019].
Why?	<i>Explain to justify / to control / to improve / to discover</i> [Adadi and Berrada, 2018] <i>to build trust</i> [Suresh et al., 2021, Mohseni et al., 2021b].
What?	For example: <i>What did the system do? / Why did the system do P? / Why did the system not do X? / What would the system do if Y happens? / How can I get the system to do Z, given the current context?</i> [Ribera and Lapedriza, 2019]. See also [Liao et al., 2020]’s <i>question bank</i> .
When?	<i>Before / during / after the task</i> [Nourani et al., 2021]. Depends on the context: <i>human-in-the-loop / human-on-the-loop / testing the system / ex-post investigation</i> [Maxwell, 2023].
How long?	How long will the user explore the explanation? [Gajos and Mamykina, 2022, Stumpf et al., 2009].

Table 2.1: The different classifications of audiences, goals, explanation content, explanation timing and contexts presented in the XAI literature.

2022, Shin, 2021, Graaf and Malle, 2017, Lombrozo, 2006, Liao and Varshney, 2022, Danry et al., 2023], on interviews [Sun et al., 2022, Liao et al., 2023, 2020, Ehsan et al., 2021, Maltbie et al., 2021, Tsai et al., 2021, Kim et al., 2023] or participatory design [Panigutti et al., 2023a, Cheng et al., 2022, Wang et al., 2019a] to learn about users’ contexts and needs. These approaches form the starting point of the HCI disciplinary triangulation between natural science theory, artefact design, and scientific observations to design empowering explainability systems [Mackay and Fayard, 1997].

Using interviews, articles such as [Sun et al., 2022, Liao et al., 2020, Lim and Dey, 2009] give fine-grained accounts of users’ questions and motivations regarding explainability. They inform on the actual user demand for information about AI systems, in various contexts, for example AI development and debugging [Zhu et al., 2018, Krause et al., 2016, Sun et al., 2022, Kulesza et al., 2015]; ideation with AI for designers [Liao et al., 2023]; doctor assistance in healthcare [Panigutti et al., 2023a, Wang et al., 2019a, Caruana et al., 2015, Jin et al., 2020, Jacobs et al., 2021]; or pretrial risk assessment [Yacoby et al., 2022]. For example, Ehsan et al. [2021] interviewed 29 AI users and practitioners to learn about the socio-organizational context of XAI-aided decision making, a perspective they call "Social Transparency". Sun et al. [2022] conducted workshops with 43 software engineers to explore their explainability needs when using generative AI for code. Maltbie et al. [2021] conducted stakeholder interviews to implement XAI in the public sector for sewer overflow predictions.

Some studies have also summarized the wide range of questions that users can have on AI systems [Liao et al., 2020, Lim and Dey, 2009]. Liao et al. [2020] employed card-sorting exercises to encourage participants to sort the most important questions they had.

Scenario-based design [Carroll, 1997], in which participants are engaged in a scenario to elicit their feedback, has often been used to understand explainability users in context [Cirqueira et al., 2020, Sun et al., 2022, Wolf, 2019, Liao et al., 2023].

Another challenge that HCI researchers are tackling is the capture of users' mental states when they are interacting with AI systems. Work in the social sciences has highlighted the importance of (1) identifying the specific knowledge gap and the foil that the explainee is trying to address, and (2) keeping track of what the explainee already knows, as seen in Section 2.3. This allows for more relevant explanations. Some efforts to capture dynamically users' specific questions and mental representations of AI systems and domains are starting to emerge in the explainability literature. This what several currents known as conversational XAI [Ehsan et al., 2019, Grimes et al., 2021, Madumal et al., 2019, Weitz et al., 2021, Hernandez-Bocanegra and Ziegler, 2021], interactive XAI [Chromik et al., 2021, Ooge et al., 2022] and interactive ML [Teso et al., 2023, Amershi et al., 2014, Guo et al., 2022] are working towards. Early AI systems used human-like communication processes to provide explanations in the form of dialogues and conversations in natural language [Abdul et al., 2018]. In 1986, for example, Winograd and Flores [1987] stressed the need for explanation systems to reflect the user's mental representation of the domain [Broniatowski, 2021].

2.4.4 Designing explainability systems

Design methods from user experience research such as card sorting, participatory design or scenario-based design have sometimes been used to ideate and conceive explainability interfaces.

Low-fidelity prototypes with conceptual artefacts as test explanations have sometimes been proposed to build and test ideas quickly. These were often put in context, through scenario-based design [Cirqueira et al., 2020, Sun et al., 2022, Wolf, 2019, Liao et al., 2023, Tsai et al., 2021].

Higher fidelity prototypes in which an explainable technique (XAI) is programmed were used more frequently [Kulesza et al., 2015, Cheng et al., 2019, Krause et al., 2016, Chromik and Butz, 2021, Panigutti et al., 2023a, Wang et al., 2019a, Springer and Whittaker, 2019]. Kulesza et al. [2015] drew on existing literature and design principles to develop their prototype. Panigutti et al. [2023a] and Wang et al. [2019a] used co-design methods to involve end-users in designing solutions [Rogers et al., 2023, International Organization for Standardization (ISO)]. Wang et al. [2019a] sketched initial visualisation prototypes, which they improved through five iterations with clinician participants. Then, in 14 co-design sessions with a clinician participant, they extracted key design implications for explainability interfaces in the medical domain, such as *"supporting access to source and situational data"* or *"supporting forward (data-driven) reasoning by showing feature values and attributions before class attribution to avoid confirmation bias"*. Panigutti et al. [2023a] redesigned their explainability interface based on users' feedback on an initial prototype and then relied on heuristic evaluation⁹ to test the usability of the new interface. The

⁹Heuristic evaluation is a method for identifying problems in a user interface (UI), which involves a team of evaluators judging it according to a set of usability guidelines [Nielsen, 1992].

in a “you’ll know it when you see it” fashion, as this is prone to confirmation bias and unscientific practices. They introduced three different approaches to evaluate explanations. These approaches are *functionally-grounded*, *human-grounded* and *application-grounded*, from less to more domain-specific and costly. **Functionally-grounded evaluation**, also referred to as algorithm-centered evaluation [Ooge, 2023], is a method that does not involve human participation and relies on statistical metrics to quantify the effectiveness of an explanation. Several common metrics used in this approach include stability, robustness, consistency, sparsity, discriminativeness, and computational efficiency [Afchar et al., 2022]. **Human-grounded evaluation** requires human participants to rate explanations along various criteria, or complete tasks such as simulating a model’s prediction for a given input after seeing explanations of the model’s behavior. Human-centred evaluation does not involve real users in specific applications. Instead, it usually involves artificial tasks that enable the testing of explanations with a large panel of human participants. Lastly, **application-grounded evaluation** consists in testing explanations in real-world settings, with real users. Users are instructed to engage with explanations and subsequently provide feedback on their subjective experience. For example, participants rate their level of satisfaction, subjective trust or perceived utility of the explanations. They may also answer questions that allow researchers to determine the amount of knowledge they gained or the extent to which they relied on the AI [Poursabzi-Sangdeh et al., 2019].

	Co-12 Property	Description
Content	Correctness	Describes how faithful the explanation is w.r.t. the black box. Key idea: Nothing but the truth
	Completeness	Describes how much of the black box behavior is described in the explanation. Key idea: The whole truth
	Consistency	Describes how deterministic and implementation-invariant the explanation method is. Key idea: Identical inputs should have identical explanations
	Continuity	Describes how continuous and generalizable the explanation function is. Key idea: Similar inputs should have similar explanations
	Contrastivity	Describes how discriminative the explanation is w.r.t. other events or targets. Key idea: Answers “why not?” or “what if?” questions
	Covariate complexity	Describes how complex the (interactions of) features in the explanation are. Key idea: Human-understandable concepts in the explanation
Presentation	Compactness	Describes the size of the explanation. Key idea: Less is more
	Composition	Describes the presentation format and organization of the explanation. Key idea: How something is explained
	Confidence	Describes the presence and accuracy of probability information in the explanation. Key idea: Confidence measure of the explanation or model output
User	Context	Describes how relevant the explanation is to the user and their needs. Key idea: How much does the explanation matter in practice?
	Coherence	Describes how accordant the explanation is with prior knowledge and beliefs. Key idea: Plausibility or reasonableness to users
	Controllability	Describes how interactive or controllable an explanation is for a user. Key idea: Can the user influence the explanation?

Figure 2.9: The 12 Explanation quality properties proposed by [Nauta et al., 2023].

Hoffman et al. [2019] suggested that the three tests of satisfaction, understanding and performance are key to measuring the “goodness” of explanation. The paper also presents an Explanation Satisfaction Scale

and summarizes the different ways to elicit users' understanding and the different approaches to measure performance of the (X)AI + human team at conducting the tasks for which the technology is designed. Additionally, it provides a checklist to measure users' curiosity and trust measurement scales. Vereschak et al. [2021] conducted a thorough review of trust measurement for explainable AI. Holzinger et al. [2020] proposed a System Causability Scale, similar to the System Usability Scale [Jordan et al., 1996], to determine whether an explanation is suited to an intended purpose. More recently, Nauta et al. [2023] reviewed explanation evaluation strategies in XAI and presented a grid of twelve properties for assessing explanations. Three of these properties require input from users: context, coherence, and controllability. The other properties pertain to explanation content and presentation, as shown in Figure 2.9.

2.5 *Explainability in Law: dreaming in color?*

Law has long recognized the need to impose information disclosure on certain, generally powerful, actors. Justice Louis Brandeis's saying that "Sunlight is the best disinfectant" has inspired transparency obligations in a broad range of fields [Schauer, 2011, Lee, 2017]. Law and economics scholars have traced the need for information disclosure to various market failures, such as information asymmetries and monopoly [Daniels et al., 2019, Wolfe, 2013]. It is no surprise therefore that information disclosure obligations have found their way into legislation on algorithmic transparency and explainability.

2.5.1 *Legal requirements for algorithmic explainability*

In 2016, legal scholars started to analyze the legal foundations of explainability for machine learning models [Kroll et al., 2016, Selbst and Barocas, 2018, Wachter et al., 2017]. Legal scholars pointed to preexisting obligations to explain algorithmic decisions, which existed well before the advent of deep learning models and before the term "explainable AI" became fashionable. These obligations were found for example in the 1995 European Data Protection Directive and in the US Fair Credit Reporting Act of 1970.

Today explainability can be found, with different names, in numerous EU legal texts that do not specifically target AI.

The GDPR (General Data Protection Regulation 2016/679, GDPR) [European Parliament and Council, 2016], requires disclosure of "*meaningful information about the logic involved*" (articles 13-15) in fully automated decisions. The GDPR provisions apply "when the decisions (i) involve the processing of personal data, (ii) are based solely on an automated processing of data and (iii) produce legal or significant effects on the recipient of the decision" [Bibal et al., 2021]. According to Maxwell and Dumas [2023], the GDPR requirements correspond to both local and global explainability.

Several explainability obligations concern platform regulation, which aims at protecting consumers and business users of platforms. The Digital Services Act ("DSA") requires disclosure of "*meaningful information directly and easily accessible [...] about the main parameters*" of recommender systems (art. 26) and more generally of the "*reasons for the relative importance of those parameters*" (art. 27) [European Parliament and Council, 2022]. As stated by Maxwell and Dumas [2023], the decision of whether the given "reasons" should faithfully and logically represent the actual system behavior will be left to regulators and the CJEU¹⁰. The Platform to Business (P2B) Regulation [European Parliament and Council, 2019] mandates that business users of platforms have access to information on algorithmic parameters to allow for an "*adequate understanding*" of the ranking and recommendation algorithms used, and that the main parameters and their importance be justified. The Proposed platform workers' directive contains similar provisions to disclose the main parameters used by algorithmic systems and their relative importance. Addition-

¹⁰ Court of Justice of the European Union

ally, consumer protection law also has provisions regarding explanations of recommender systems in online marketplaces. It notably imposes to show “the main parameters determining ranking [...] of offers presented to the consumer as result of the search query and the relative importance of those parameters as opposed to other parameters” (new art. 6(a) of Directive 2011/83 on Consumer Rights).

Bibal et al. [2021] also emphasize that explainability requirements are stronger in the public sector. Any decision made by a public authority, such as an administration or a judge, must always be justified and reasons for the decision must be clarified and explained. When the administrative decision-making process is automated, further explainability requirements may be necessary. French administrative law is among the most demanding frameworks, requiring that the person subject to the decision be able to request the parameters used in the process and their weighting (art. R. 311-3-1-2 of the French Code on the relationships between the public and the administration) [Maxwell and Dumas, 2023].

The above section does not provide a comprehensive list of all the provisions for explainability in legal texts or decisions. Rather, it gives a brief overview of the ways in which explainability may be provided in law. Below we extend the discussion by focusing on two of the most cross-sectorial legal foundations for explainability: the upcoming AI Act¹¹ and human rights case law.

Explainability and the AI Act

In December 2023, the EU reached an agreement on the text of the AI Act, which aims to harmonise regulation on AI systems and make the EU the first region in the world to do so. The text promotes a regulatory approach based on the level of risk that AI systems pose to fundamental rights. It sets out different obligations depending on whether the AI application falls into one of these four risk categories:

- **Unacceptable risk:** this includes systems that comprise manipulation, exploitation, social scoring, or biometric identification of people. These AI applications will be strictly prohibited, with very limited exceptions.
- **High-risk:** AI applications in critical sectors such as transport, education, employment and health or law enforcement are among the areas concerned. For example, AI systems used to evaluate individuals’ creditworthiness are considered as high-risk. AI applications that fall within products already regulated by EU law, such as an AI-based diagnostic tool used in healthcare, are considered high-risk. High-risk AI system suppliers will have to carry out a prior conformity assessment and satisfy other requirements to ensure the safety of their AI systems before putting them into service in the EU. Suppliers are also bound to transparency requirements to provide information on high-risk AI systems for all stakeholders.
- **Limited risk:** Systems with low risk should meet basic transparency

¹¹ As this thesis was written between September 2023 and January 2024, the final text of the AI Act had not been published yet. We therefore relied on a near final draft version in the sections below.

requirements, such as informing users that they are interacting with an AI, allowing them to make informed decisions.

- **General purpose and generative AI:** The initial proposal of the European commission did not account for "general-purpose AI models" or foundation models. The trilogue discussions in late 2023 have integrated generative AI regulation in an entirely separate risk class. Inside this class, generative AI models that are used for research and development and not used in the EU market are exempt from obligations. Other generative AI models will have to comply to transparency requirements such as disclosing that content was generated by an AI, preventing the model to produce illegal content, and disclosing copyrighted data used for training. In addition, models that may pose systemic risk, such as the latest GPT-4, will have to undergo more thorough risk evaluations. Classified in this category are models for which the compute power exceeds 10^{25} FLOPS.

It should be noted that the rules are designed to be evolutionary: the definition of AI or the quantitative criteria for considering a model to represent systemic risk could easily change.

Classification of the AI applications studied in this dissertation.

In Chapter 5 and 6, we consider two applications of AI systems which may be considered high-risk under the AI Act.

The first one involves using **an AI system to provide an online recommendation for a life insurance plan** that matches a user's financial situation. These systems are called "robo-advisors". It is clearly covered by Annex III of the draft AI Act which lists high-risk AI applications: "*AI systems intended to be used for risk assessment and pricing in relation to natural persons in the case of life and health insurance.*"¹²

¹² Annex III, paragraph 6.

The second application of AI we consider is **the detection of money laundering and terrorist financing**. Anti-money laundering and counter-terrorism financing (AML-CFT) systems are implemented by banks, which are required to report suspicions of money laundering or terrorist financing in their customer base to financial intelligence units (FIUs). In turn, FIUs investigate these suspicions in order to refer serious ones to law enforcement authorities. It is unclear if AI systems used in AML-CFT systems can be considered as "high risk" under the AI Act. Some scholars have interpreted it could be the case [Pavlidis, 2023], considering a former point in the Commission proposal, which has been removed in the most recent AI Act draft. Nevertheless, Annex III, point 7(e) and specifically 7(f) could be interpreted as applying to AI systems in AML-CFT: "*AI systems intended to be used by law enforcement authorities or on their behalf or by Union institutions, agencies, offices or bodies in support of law enforcement authorities for profiling of natural persons as referred to in Article 3(4) of Directive (EU) 2016/680 in the course of detection, investigation or prosecution of criminal offences.*" However, recital 37 foresees an exception for "*AI systems used for the purpose of detecting financial fraud*". In light of these provisions, it is more likely that AI systems used by banks to enhance ML/TF detection would not be considered high risk. However, one could

argue that money laundering, terrorism financing and financial fraud are distinct concepts [Unger and Busuioc, 2007]. The systems put in place to prevent money laundering target a larger scope of criminal offenses than fraud, including, for example, human and drug trafficking.

The role of explainability in the AI Act.

Panigutti et al. [2023b] highlights that the AI Act does not mandate a requirement for explainable AI, but rather aims to achieve trustworthy AI through the pillars of transparency and human oversight. The authors consider, however, that implementing such measures may be done through use of explainable AI. As Maxwell and Dumas [2023] notes, humans in charge of oversight should be "able to correctly interpret the high-risk AI system's output, taking into account in particular the characteristics of the system and the interpretation tools and methods available"¹³ For Maxwell and Dumas [2023], this indirectly suggests the need for local explanations.

¹³ art. 14-4(c) of the Commission's proposal for the AI Act.

Explainability in human rights case law

Decisions of the Court of Justice of the European Union (CJEU)¹⁴ also inform us on the need for explainability with regard to fundamental rights protected by the Charter [Maxwell and Dumas, 2023]. Maxwell and Dumas [2023] unpack those explainability requirements. In the *Ligue des droits humains v. Council of Ministers* case, the CJEU said that AI systems which decisions can lead to serious consequences should rely on "pre-determined models and criteria", therefore calling for global explainability and excluding the use of machine learning. Furthermore, high-risk AI systems, such as those used for terrorism detection, should provide explainability to enable human operators to evaluate the generated alerts. The CJEU also considers that local explainability enables contestability, which falls within an individual's due process rights.

¹⁴ CJEU, 6 October 2020, *La Quadrature du Net*, joined cases C-511/18, C-512/18 and C-520/18; CJEU, 21 June 2022, *Ligue des droits humains v. Council of Ministers*, Case C-817/19

2.5.2 Legal objectives for explainability

The objectives of regulation are intertwined with economic goals to correct market failures [Levine and Forrence, 1990] such as information asymmetry, customer abuse, trade secrets, economic crime or distrust in the economy and institutions. These regulatory ambitions are reflected in the purposes of legal requirements for explainability, which are outlined by Maxwell and Dumas [2023]. Further, the appeal towards explanations in legal texts can be attributed to the notion of reason-giving in law, as argued by Rozen et al. [2023], which pursues specific objectives. Below, we summarize the explainability purposes of explainability and reason-giving as presented by Rozen et al. [2023] and Maxwell and Dumas [2023]:

1. **User empowerment.** Requirements for global explanations enable individual or business users to access minimal information to understand algorithmic recommendations and preserve their agency. This reflects a regulatory concern to correct information asymmetries and protect

consumers. This objective corresponds to acknowledging the human agency of the decision subject as described in [Lombrozo, 2012].

2. ***Evaluation and quality of individual decisions.*** [Maxwell and Dumas, 2023] contends that providing local explanations may be necessary to allow for effective human oversight of individual decisions, which is a fundamental right protected by the EU Charter. This aligns with the primary purpose of reason-giving in law, which is to ensure fair and just decisions [Rozen et al., 2023].
3. ***Contestability and due process.*** Provisions for local explainability aim to enable individuals to challenge decisions. This stems from regulatory goals to protect individuals' fundamental rights to quality decisions concerning them and due process of administrative decisions. For example, Margot Kaminski and Urban [2021] discusses what an individual right to contest algorithmic decision should look like, building on the United States' tradition of due process theory.
4. ***Control over system performance.*** Explainability is also needed to check that systems used to pursue general interest objectives are sufficiently efficient, such as AI-based anti-money laundering systems, for example.
5. ***Accountability and legitimacy of decision makers.*** Additionally, legal requirements for explainability may arise from the need to preserve transparency in public administration [Maxwell and Dumas, 2023], in order to preserve public trust in institutions. This is in line with Rozen et al. [2023]'s view that reason-giving serve the purpose of promoting compliance and legitimacy of deciding bodies. They quote Jerry Mashaw who asserts that "*the authority of all law relies on a set of complex reasons for believing that it should be authoritative*" [Mashaw, 2001]. In this context, explanations serve as accountability mechanisms in socio-techno-legal contexts in which human deciders are concerned with reputational risks, peers' approval or other incentives to make the "right" decision [Rozen et al., 2023].

2.5.3 *Is explainability the best disinfectant?*

"Sunlight is said to be the best of disinfectants; electric light is the most efficient policeman"

Louis Brandeis, 1913

Explainability for decision-subjects empowerment and contestability.

Returning to Louis Brandeis' famous saying, transparency can be seen as a remedy to corruption and illegitimacy in politics and society. However, there are opposing views and nuances to consider. Here, the judge takes "electric light" as a metaphor for a "technology of transparency" that enables effective oversight and enforcement [Obar, 2020]. In explainability, it amounts to giving access to explanations of algorithmic decisions to decision-subjects and citizens, as a way to achieve greater accountability and trustworthiness of AI systems. However, Wachter et al. [2017] state: *"the feasibility and practical requirements to offer explanations to data subjects remain unclear."* In fact, many legal scholars have criticised Brandeis' vision as overly simplistic, advancing that it may represent an ideal, but an unattainable one [Lippmann, 1993]. Jonathan Obar [2020] argues that advocating transparency is one thing, but achieving "meaningful forms of transparency" is more difficult. Taking the example of consent to personal data practices, the author observes that the self-governance fallacy is deeply ingrained in the occidental democratic approach. Indeed, as Pasquale [2015] puts it, *"discovering problems in Big Data should not be a burden we expect individuals to solve on their own"*. Obar [2020] therefore asserts the need to recognise human limitations and to move the discussion beyond on access to information, and rather towards what happens afterwards, raising questions such as how do we effectively communicate information to end-users and how do we support engagement with the content of the message? and is that even realistic?

Therefore, focusing on explanation design, representation and communication could provide some answers to the propensity of explainability for lay users to meet legal objectives. In the context of GDPR requirements, Wachter et al. [2017] defend that there should be more efforts to *"determine whether and how explanations can and should be offered to data subjects (or proxies thereof) with differing levels of expertise and interests."* We explore in Chapter 5 this tension between the capabilities and needs of decision-subjects on the one hand, and the ideal of appropriate trust calibration, on the other. Specifically, we explore this in the context of AI-based recommendations for life-insurance plans, where non-expert end-users should be given clear, concise and non-misleading information in order to make an informed choice.

Explainability for decision quality, due process and accountability.

Even for audiences other than decision-subjects and citizens, there is growing scepticism from law scholars about whether explainability can achieve legal objectives. Rozen et al. [2023] are rather sceptical about explainability's propensity to contribute to the objectives of reason-giving

in law.

First, for the authors, explainability cannot contribute to restraining and slowing down human judgement for "a better and more just decision" because it is not humans but machines that are making decisions. This objective for reason-giving in law relies on the human nature and our capacity to "feel" accountable, which is not applicable to machines in the XAI context.

Second, Rozen et al. [2023] emphasizes the problem of unreliable explainability methods and the difficulty it creates to meet due process requirements. Wachter et al. [2017], also highlight that leveraging algorithmic audits is critical to *"provide an evidence trail for providing explanations of automated decisions."* We study the role of explanations for AI auditing in Chapter 6, where we describe the approaches and needs of regulatory supervisors for auditing AI-based anti-money laundering systems.

Third, while explanations provide "clues" and approximations about the model behavior, they require human deduction skills to be interpreted, and humans can potentially be manipulated in that process. This makes it harder to challenge decisions and facilitate due process rights relying solely on explanations. We explore this in detail in Chapter 3, where we uncover the different human biases at play in explainability interpretation.

Finally, Rozen et al. [2023] concede that explainability can play a role in strengthening the accountability and authority of decision-makers. We also explore this aspect in chapter 5 by considering the role of explainability in strengthening the accountability of life insurance providers, and in Chapter 6, where explanations are used as means to increase accountability and auditability of banks regarding their AI-based anti-money laundering systems.

PART I

*Calibrating trust in explainable
AI: common pitfalls and the
promise of interactivity*

Chapter 3: Trust, overtrust, distrust in explainable AI: a cognitive approach presents a review of the cognitive biases in explainable AI literature. This chapter builds on an article that was published as a conference paper:

"How Cognitive Biases Affect XAI-Assisted Decision-Making: A Systematic Review", Astrid Bertrand, Rafik Belloum, James R. Eagan, Winston Maxwell, Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22), Oxford, UK, 2022 <https://doi.org/10.1145/3514094.3534164>.

This thesis deepens the analysis presented in the conference paper. As the first author, I delineated the motivation and research questions. I led the review process and was helped by the second author to classify and analyze the papers. I wrote most of the paper, specifically the findings and discussion. The methods, results, and text were discussed with all three co-authors.

Chapter 4: Towards "human-like" explanations: the promise of interactivity presents a detailed scoping review on interactive explainable AI. This chapter builds on an article that was published as a conference paper:

"On Selective, Mutable and Dialogic XAI: A Review of What Users Say about Different Types of Interactive Explanations", Astrid Bertrand, Tiphaine Viard, Rafik Belloum, James R. Eagan, Winston Maxwell, Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), Hambourg, Germany, 2023 <https://doi.org/10.1145/3544548.3581314>.

As the first author, I delineated the motivation and research questions. I led the review process and was helped by the second and third authors to classify and analyze the papers. I wrote most of the paper, specifically the findings and discussion. The methods, results, and text were discussed with all co-authors.

Chapter 3

Trust, overtrust, distrust in explainable AI: a cognitive approach

"Automated decision aids are designed to reduce human error, but actually can cause new errors in the operation of a system if not designed with human cognitive limitations in mind".

Cummings [2004]

AT THE HEART OF human-computer interaction research is the search for optimal collaboration between humans and machines. Trust plays a significant role in this collaborative relationship, as it determines the extent to which users will use the machine's advice when faced with complex or uncertain situations [Culley and Madhavan, 2013, Lee and See, 2004]. We therefore begin our research with a characterization of the cognitive challenges to trust explainable AI systems. We review of the cognitive ways in which people trust, but also overtrust, distrust or misuse explanations by searching the literature in explainable AI. We highlight important individual and contextual factors in the trust calibration process. This allows us to emphasise the relevance of human-centred approaches to explainability design.

Section 3.1 presents the motivation and research questions for the survey presented in this Chapter. We build on HCI research regarding human biases when working with automation, as well as on work in sociology and philosophy of science regarding cognitive aspects of explanations. Section 3.2 describes this prior research. Section 3.3 develops the methodology used for the review. Section 3.4 presents the results, including the cognitive mechanisms explanations should adapt to, the way explanations can be misused and disused through users' cognitive biases, or misevaluated in user studies. We also describe the bias mitigation strategies identified in the explainability literature. Finally, Section 3.5 discusses avenues in explainability research to take into account identified pitfalls.

3.1 *Motivation and research questions*

Correctly calibrating trust in AI decisions and systems may be paved with important cognitive challenges, including Automation-Induced Complacency (AIC) [Parasuraman et al., 1993], and possibly other biases.

While there are growing efforts from researchers [Green and Chen, 2019, Mittelstadt et al., 2019, Rastogi et al., 2020] to tie cognitive science literature to a mostly technical explainability field, more research is needed to identify what kind of cognitive biases and heuristics are involved in the explanation process, and whether and how to leverage people’s heuristics to improve XAI systems. Several studies exist that shed light on cognitive mechanisms leading issues when interpreting explanations of AI systems [Chromik and Butz, 2021, Wang et al., 2019a]. However, the literature lacks a comprehensive review of the efforts made so far on this front in the explainable AI field. In this chapter, we focus on cognitive biases in order to pin down the cognitive challenges to fostering appropriate trust in explainable AI.

To the best of our knowledge, there is not yet a comprehensive review of how cognitive biases have been accounted for in the explainability literature. A analysis like the one we present appears necessary to summarize findings on how cognitive biases interfere with explanations, how to address them, and to highlight promising directions concerning the integration of cognitive processes in XAI systems.

In this work, we consider cognitive biases not only in terms of “errors” (e.g., automation bias that leads to inappropriate trust in AI modes) but also as the cognitive constraints that are inherent in the human explanation process.

We analyze how the field of XAI has been dealing with human cognitive biases and constraints, and we discuss promising mitigation strategies and research directions to support human critical thinking. To this end, we conducted a scoping review of 38 papers, based on a systematic search methodology, and guided by the following five research questions:

- RQ1:** *What cognitive biases have been studied in the explainability literature?*
- RQ2:** *In which contexts (e.g., explainability method, human expertise, tasks type) do these cognitive biases arise?*
- RQ3:** *How to adapt to human cognitive architecture to improve explainable AI systems?*
- RQ4:** *What evaluation methods have been used to detect cognitive biases (specific to each bias)?*
- RQ5:** *What are the stated future research directions and challenges identified by the scientific community?*

3.2 Background

3.2.1 Trust in automation

Decades of research at the intersection of psychology and HCI research highlight important and pernicious challenges to appropriately calibrate trust in automated intelligent decision support systems [Parasuraman and Riley, 1997, Bailey and Scerbo, 2007, Lee and Moray, 1992, Wickens et al., 2009, Gawronski, 2004, Cummings, 2004]. This literature emerged from the study of complex systems in critical environments, specifically the aeronautics in the 1990s. The analyses of plane crashes, such as the 1996 accident [National Transportation Safety Board, 2000], shed light on difficulties for pilots in understanding system warnings, detecting automation errors or monitoring highly reliable systems, leading to catastrophic consequences [Billings, 1996].

Definition

Complacency. Parasuraman et al. [1993] described the phenomenon of *Automation-Induced Complacency (AIC)*, which is a state of "low suspicion" by the human operator when the automation performs a task for them, also defined as "self-satisfaction" resulting in non-vigilance".

The term "complacency" is necessary because it encompasses constructs broader than vigilance failure, boredom, or workload issues. Complacency represents a unique attitude, and complacency and boredom are not connected [Parasuraman et al., 1993].

A related notion in the literature is automation bias. According to Cummings [2004]:

Definition

Automation bias. Automation bias "occurs when a human decision maker disregards or does not search for contradictory information in light of a computer-generated solution which is accepted as correct" [Cummings, 2004].

Complacency and automation bias have often been discussed as separate concepts in the literature [Parasuraman and Manzey, 2010]. On the one hand, complacency involves a lack of attention, predominantly observed in conditions of multitasking, and high automation reliability. On the other hand, automation bias is seen as a tendency to overtrust decision-support systems. By noting these differences, we can see that they are due to variances in the observation of these concepts. However, ultimately both notions result in the same underlying problem. If we take Ferrario et al. [2020]'s definition of trust which involves the lack of monitoring, automation bias becomes very similar to complacency. In fact, Parasuraman and Manzey [2010] argued that "automation-induced complacency and automation bias represent closely linked theoretical concepts that show considerable overlap with respect to the underlying processes". Therefore, for simplicity, we will consider the two terms as synonymous in the remainder of the dissertation.

Additionally, we summarize below significant factors determining trust in automation, building on Culley and Madhavan [2013]’s review. These factors include: variability of system reliability, operator cognitive load (e.g. multitasking), alarm threshold, severity of the consequences of failure or trust in the system designer.

Human operators are not well suited to monitoring infrequent and unanticipated problems in complex systems, particularly when the system is highly reliable and the operator is multitasking [Bailey and Scerbo, 2007, Parasuraman et al., 1993]. In general, system reliability and performance have a great effect on operator trust [Bailey and Scerbo, 2007]. AIC occurs over time, after a period of familiarisation with automation [Molloy and Parasuraman, 1996]. Varying system reliability eliminates complacency effects [Parasuraman et al., 1993, Bailey and Scerbo, 2007]¹. Wickens et al. [2009] investigated the “cry wolf effect”, whereby low alarm thresholds and a surplus of alarms result in an operator’s distrust and disregard of the alarm system, potentially leading to the neglect of true alerts.

When systems make mistakes, the loss of trust is proportional to the severity of the consequences of the failure [Culley and Madhavan, 2013]. However, *difficult* or *near* misses can result in a lower loss of confidence [Madhavan et al., 2006].

Parasuraman and Riley [1997] also emphasised the importance of trust in the human designer of the system as a key factor in calibrating trust in automation.

Finally, Lee and See [2004] introduced the notions of *resolution* and *specificity* of trust. Resolution is the ability to adjust one’s confidence in proportion to changes in the system’s capabilities. A person with a low confidence resolution will only slightly change their confidence in a system that has undergone major changes to its capabilities. Specificity refers to the ability to calibrate one’s trust in all the different system’s distinctive components.

3.2.2 Trust in automation by AI systems

Over 40 years after a first research wave on automation provided by intelligent decision-support systems, with the difference that systems are even more complex and opaque. Findings from early research on trust in automation appear more topical today than ever [Zerilli et al., 2019, Cummings, 2004].

Glikson and Woolley [2020] highlight the differences between the traditional automation that was the subject of early studies on complacency and automation of decisions by modern AI systems. They define traditional automation as “systems that perform repetitive and monotonic tasks that were previously performed by humans” [Parasuraman and Riley, 1997, Glikson and Woolley, 2020]. These systems are deterministic and their behavior is known and fully pre-programmed. On the contrary, machine learning models execute tasks significantly differently from the human approach, primarily because of their probabilistic nature and ability to learn from large data.

¹ During the 1990s, Airbus planes were the most automated commercial planes in operation. To prevent automation bias, pilots were warned against becoming excessively dependent during training. Following an Airbus plane crash in 1992, French airlines implemented a policy requiring pilots to periodically take manual control of automated systems.

Glikson and Woolley [2020] then review studies on trust in AI and reveal six factors enabling cognitive and emotional trust. These are tangibility, transparency, reliability, task characteristics, immediacy behaviors and anthropomorphism. Tangibility refers to the different forms that AI can embody, from physical presence as in the case of robots, to virtual agents or bots or to AI embedded in computers. Humans tend to trust more AI systems that are more tangible in this order: physical > virtual > embedded. Transparency and explanations of AI systems tend to increase trust. Low levels of reliability significantly reduce trust, and it is difficult and time-consuming to regain it. For tasks that require data analysis, AI is trusted more while for tasks that require social skills, AI is trusted less than humans. Immediacy behaviors refer to personalization, interactivity, adaptiveness and responsiveness, which are usually associated with increasing trust.

Furthermore, Stanton and Jensen [2021] identify other factors that affect human trust in AI, namely usability of AI systems (*i.e.* the user experience), and the technical characteristics identified by HLEG's definition of trustworthy AI (accuracy, reliability, security, explainability, privacy...)

Zerilli et al. [2019] also ties the research on complacency and trust in automation by intelligent systems with the more recent trends in automation by machine learning and AI systems.

Zerilli et al. [2019] focus on the "control problem", which is broader than the issue of trust in automation. Control here pertains to the capacity to diagnose and address faults or issues as they arise in real-time, as well as to proactively address future issues. Zerilli et al. decompose the control issue into three main sub-problems: *the capacity, attentional and attitudinal problems*. *The capacity problem* refers to the lack of processing power of human architecture compared to computer processing, that make them inherently unable to monitor in real time a task they cannot do themselves [Bainbridge, 1983]. Zerilli et al. [2019] argue how this becomes particularly salient in the age of deep learning, where even software engineers cannot fully understand the "multi-vector logic" of a neural network. *The attentional problem* refers to humans' limits in term of attention over time. It refers to studies on "vigilance" that point to the cognitive impossibility for humans to maintain effective visual attention on an interface on which little happens [Bainbridge, 1983]. Finally, the *attitudinal problem* refers to humans' tendency to believe that the system is reliable enough to be left alone. It therefore refers precisely to a trust calibration issue, and to the observations in complacency and automation bias studies.

Overall, research in psychology and HCI has shown that humans are at a severe disadvantage to occupy monitoring functions of complex and autonomous systems [Bainbridge, 1983]. Zerilli et al. [2019] claim that there are no reason to believe that the human tendencies observed with early automation, that result from million years of evolution, would not manifest with machine learning systems. Does this mean that human and AI collaboration is doomed? Zerilli et al. [2019] add nuance to that view. First, some AI systems show impressive levels of performance that exceed those of well-trained humans, making it inconsequential that hu-

mans cannot monitor machine decisions. For example, there are AI systems which can detect Alzheimer’s disease with over 80% accuracy ten years before the appearance of the first symptoms. Depriving ourselves of the capabilities of this algorithm for reasons of human control would be a major opportunity cost for healthcare. Second, Zerilli et al. [2019] argue that in acknowledging unavoidable human biases with automation, we can work towards complementary and dynamic allocation of tasks between humans and AIs.

In this context, explainability represents an additional way out of what seems like a dead end for appropriate human oversight and trust calibration. It promises to remedy to the capacity problem by producing human-intelligible explanations and potentially to the attitudinal problem by enabling correct trust calibration. However, Glikson and Woolley [2020] and Stanton and Jensen [2021] point out studies that showed that transparency overall reinforce trust. The effects of explainability for human cognitive architectures and cognitive trust mechanisms are still unclear.

This chapter focuses on the concept of cognitive bias to examine how explanations can either bolster or undermine trust in AI.

3.2.3 *Explanations are biased and (maybe) biasing*

“In the context of explanation and revision, the strength of causal reasoning and the weakness of diagnostic reasoning are manifest in the great ease with which people construct causal accounts for outcomes which they could not predict”.

[Kahneman et al., 1982]

In theory, explainability ought to serve as an aid for humans to regain control of AI black-boxes, restore their autonomy in decision-making with AI, and prevent their errors like complacency or automation bias. Naturally, reality is not so simple. On the contrary, some results highlight the harmful potential of explanations to amplify automation biases in high-stakes settings [Jacovi et al., 2021, Eiband et al., 2021, Wang et al., 2019a]. These findings are in line with previous research in the context of intelligent decision support systems that show that automation decision aids could cause new errors instead of reducing them [Cummings, 2004, Madhavan et al., 2006]. In fact, human cognitive architecture is not something that can be “fixed” [Lindström et al., 2022]. However, it is a key element for technology designers to consider [Cummings, 2004].

In this section, we review the cognitive processes involved in explanation and the way they are inherently biased, which is not necessarily “bad” per se.

In the 1980s, Amos Tversky and Daniel Kahneman [Kahneman et al., 1982] introduced the concept of cognitive bias as:

Definition

Cognitive biases. *“Systematic error in judgment and decision-making common to all human beings which can be due to cognitive limitations, motivational factors, and/or adaptations to natural environments.”*

In 2011, Kahneman developed the dual-process theory [Kahneman, 2011], in which he described two systems that illustrate the way we think. “System 1” reflects our fast, intuitive and emotional thinking which often leads us to make errors. “System 2” is more deliberative, logical, but also requires more effort to activate.

System 1, and cognitive biases do not necessarily have bad consequences or results: they have been developed over the course of our evolution to help us think faster, interact better with our peers or keep us safe [Kahneman, 2011]. Kahneman also describes the extraordinary abilities that result from our System 1. These biases should be seen as constraints on the problem of explainability, as integral aspects of our human nature.

As seen in Section 2.3, cognitive biases and social expectations are present when people evaluate and generate explanations.

Specifically, people select causes in a biased way by paying more attention to causes that have specific characteristics [Miller, 2019]. Lombrozo [2006] talks about “the frailties of induction”. As for Pennington and Hastie [1993], an explanation is a story that coherently puts all the pieces of evidence together to give them causal sense. The produced story is subjective, as it depends on the explainer’s world knowledge about similar events, or even knowledge about story structures. Graaf and Malle [2017] also argue that people have social expectations towards machines because they attribute human traits to them. For example, we expect AI explainers to use the framework of conversations, or tend to attribute intents to them [Graaf and Malle, 2017, Dodd and Bradshaw, 1980]. We also have cognitive biases in interpreting explanations. Although generalising from explanations is necessary and useful for learning and problem solving, it can come at the cost of over-generalising. We have previously discussed in Section 2.3.1 that generalisation is significantly linked to the similarity and diversity of the properties involved [Rehder, 2006]. Lombrozo [2006] argues that these factors can lead people to over-generalize if a novel case seems similar to the case that is explained or if the presented explanation seems to hold true in a diverse range of contexts. Specifically, explanations reinforce that effect *“by providing a more restrictive basis for generalizing from known to novel cases”*.

"Explanations can lead reasoners to override the influence of similarity. If told that herring and tuna have a disease, naive participants are more likely to extend the property to wolffish, the more similar item, than to dolphins [Shafto and Coley, 2003]. However, among fishing experts, who can generate an explanation for why the property might hold (e.g. tuna contract the disease by eating infected herring), similarity is less predictive of property extensions. Instead, properties are extended if the explanation generalizes (e.g. to dolphins, who also eat herring)."

Extract from [Lombrozo, 2006].

Although many studies have shown that XAI methods can improve users' understanding of black-box models [Lakkaraju et al., 2017, Lucic et al., 2020, Ribeiro et al., 2018], recent empirical studies have drawn attention to obstacles resulting from a mismatch between people's cognitive constraints and current XAI techniques. Specifically, there have been concerns that AI explanations can bias users and impair their decision-making process [Ghassemi et al., 2021, Kaur et al., 2020, Nourani et al., 2021]. At the root of the issue, Buchanan and Shortliffe [1985] argue, is the choice between trusting an AI recommendation or engaging in an effortful and time consuming cognitive analysis of its explanations (*i.e.* engaging System 2). People thus develop biases "*about whether and when to follow the AI suggestions*" [Buçinca et al., 2021], and AI explanations can reinforce such biases.

For example, explanations can lead to unwarranted trust in AI recommendations [Jacovi et al., 2021]. Eiband et al. [2019], show that placebo explanations elicit a similar level of trust as real explanations. Other work [Chromik et al., 2021, Fürnkranz et al., 2020, Nourani et al., 2021, Wang et al., 2019a] shows that explanations can cause reasoning errors such as backward reasoning and confirmation bias. Leveraging Kahneman's dual process theory, Kliegr et al. [2021] reviewed the effects of cognitive biases on the interpretation of AI models and provide a rich analysis of over 20 different biases. That work, however, focuses on rule-based explanations. In turn, Wang et al. [2019a] propose operational pathways between users' reasoning needs and XAI methodologies. They describe how people reason when explaining and review some common cognitive biases and the ways in which they can be mitigated. However, this work does not comprehensively cover the cognitive biases that may arise in the presence of explainable AI.

3.3 Methodology

In this section, we detail the method used for the scoping literature review and how we selected the papers for inclusion.

3.3.1 Review type

Like a systematic review [Mulrow, 1994], a scoping review [Arksey and O'Malley, 2005] includes many rigorous steps to survey the literature. Scoping reviews do not require the pre-registration of the results nor the assessment of the quality of the studies [Munn et al., 2018] as systematic reviews do, but they include similar methodological steps: the definition of research questions, a systematized search and selection process, and an analysis and reporting the results [Arksey and O'Malley, 2005]. We followed the standardized search and selection methods from the systematic review methodologies, as suggested in [Arksey and O'Malley, 2005] for scoping reviews, to ensure the replicability and transparency of our findings. In particular, we followed the steps of the Preferred Reporting Items Systematic Reviews and Meta-Analyses (PRISMA) standard [Moher et al., 2009]: paper identification, screening, eligibility evaluation and analysis procedure. In doing so, it is possible to reproduce the processes of searching, selecting, and analyzing the relevant literature. This allows us to guarantee the quality of our search and selection process, as encouraged by the PRISMA Extension for Scoping Reviews PRISMA-ScR [Tricco et al., 2018].

Scoping reviews are an appropriate survey type to examine how research is conducted on a specific topic, give a summary of the focus of the field, map key concepts, identify the types of evidence found in a field, pave the way for future systematic reviews, and identify gaps in the literature [Munn et al., 2018]. This corresponds to the objectives of study: identify, map, report and discuss the available evidence on cognitive biases in XAI.

3.3.2 Corpus creation

Our aim was to give a sense of how the XAI literature has addressed the notion of cognitive biases so far. We therefore relied on a keyword-based approach, which essentially has the advantage of ensuring transparency, reproducibility and, also, leading to more comprehensive results by sampling a wide range of work. However, it is possible that some XAI articles have addressed the notion of cognitive biases in different terms, referring to specific types of cognitive bias. However, we could not include all possible types of cognitive biases as keywords, since there are over 200. We also did not want to focus the investigation on specific types of bias in order to provide a more representative view of the different cognitive biases discussed in explainability. In addition, because we conducted our searches on ACM, IEEE, and Scopus, we may have missed other relevant work from other sources. To address these limitations, we

supplemented the keyword-search with selected papers addressing cognitive biases in XAI drawn from two authors' knowledge of the XAI field. Section 3.6 discusses the limitation of the methodology in further detail.

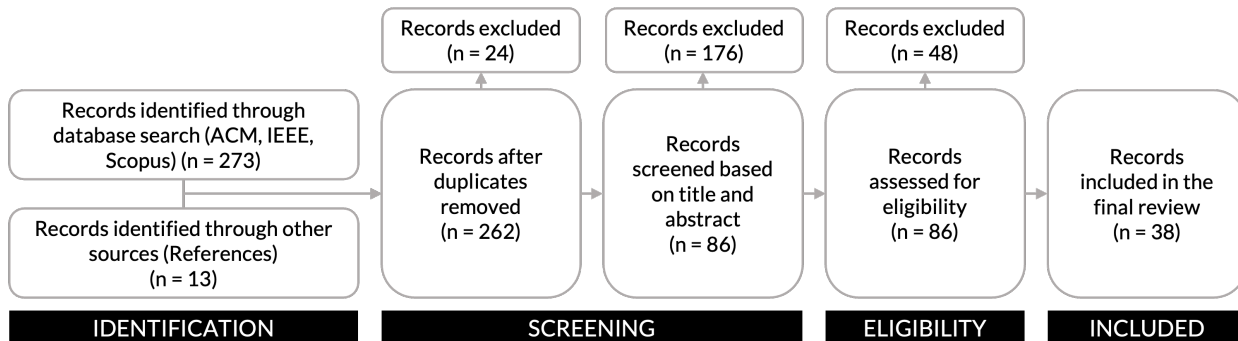


Figure 3.1: PRISMA flow diagram [Moher et al., 2009] on how the final corpus was curated (n = 38).

Keyword Match. During the identification phase, we performed a structured keyword search using the following sources: ACM, IEEE, and Scopus. Since this survey focuses on cognitive biases related to XAI, the search query was contextualized in three dimensions: AI systems, Explainability, and Cognitive biases. Drawing on the authors' background in XAI, we assigned keywords that describe each dimension. We searched for keywords representing AI systems and Explainability dimensions in the Title, Abstract, and Author Keywords fields, because we wanted to focus on papers whose main topic was XAI. For Cognitive bias keywords, we searched in the Full text of papers. The search result was filtered to include recent papers (2008 or after) since XAI is a young field of study. The search query was as follows, adapted to each database advanced search specificities (the wildcard * indicates where we retrieved plurals and different spellings):

AI systems: Abstract: (AI, artificial intelligence, machine learning, algorithm*, intelligent system*, neural network*) AND

Explainability: Abstract: (explainab*, explanation*, intelligib*, interpretab*, transparen*, XAI) AND

Cognitive biases: Full Text: (cognitive bias*, decision bias*, explanatory bias*, explanation bias*, human bias*) AND

Date: 2008 and after.

Screening and Eligibility. We considered the following inclusion and exclusion criteria. The logic followed is $(IC_1 \text{ OR } IC_2 \text{ OR } IC_3) \text{ AND } EC$.

IC₁ Cognitive biases. The paper describes cognitive biases that are involved in the field of XAI.

IC₂ Mitigation techniques. The paper describes techniques to mitigate cognitive biases involved in the XAI process.

IC₃ Measurement techniques. The paper describes ways to measure cognitive biases related to explanations.

EC Papers that do not provide primary insights on cognitive bias in XAI are excluded (e.g., a paper that does not provide enough detail on how the heuristics manifest and in what context).

Additionally, only peer-reviewed papers written in English were included. We excluded very few papers to which we did not have access. The identification phase yielded a total of 273 results: 59 papers from ACM, 64 from IEEE, 150 from Scopus, and 12 additional papers selected from the references of relevant papers or based on the authors' knowledge. The authors' names, article title, source title, and publication year of the identified records were exported to an Excel spreadsheet. A total of 261 results were obtained after eliminating 24 duplicates. In the screening stage, each paper's title and abstract was reviewed by an author based on the inclusion and exclusion criteria, and a decision was made as to whether the paper should be rejected or retained for the next phase (eligibility). 176 papers were excluded because they did not discuss cognitive biases involved in the field of XAI. A total of 85 papers were advanced to the next phase. In the eligibility stage, two of the authors read the remaining articles in full. Based on the inclusion and exclusion criteria, a decision was then made as to whether the article should proceed to the final phase. 48 articles were finally excluded at this stage because they did not sufficiently address the proposed research questions (cf. introduction). 38 articles were retained and advanced to the final phase.

Coding book. In the inclusion stage, we started the coding of the papers by having two authors extract relevant information from the papers. Except for the type of article (primary study or survey), this information essentially relates to RQ2 (see introduction). To ensure coding quality, this information was brainstormed by the authors and the research team and was drawn from related surveys of empirical studies of XAI (e.g., [Lai et al., 2021]). As such, our code book included: Cognitive bias type; Mitigation strategy; Explainability technique and format (local feature explanation, global explanation, etc.); Paper type (primary study or review); Application/domain (high-risk or low risk); AI type (shallow, deep or wizard of oz) and algorithm used (when specified); Human task type (proxy or real and description); Human expertise (lay-user, domain expert or ML expert). The full code description is presented in Table 3.1.

Corpus presentation. In the corpus of 38 papers we analyzed, 7 papers are reviews of the literature, and 31 papers are primary studies. Figure 2 illustrates the distribution of our corpus across the disciplines, showing the diversity of the subject areas. As we can see, over half of these papers are Human Computer Interaction (HCI) works, published in leading conferences (e.g., CHI and IUI). The remaining papers have also been published in leading conferences and journals directly or indirectly related to the explainability of AI systems, in the fields of AI, computer science and psychology.

Dimension	Code with examples found in the corpus
AI types	Deep learning models (<i>deep reinforcement learning, RoBERTa, Re-ID networks, BERT, CNN VGG-19, deep neural network based on GoogleNet</i>), Shallow models (<i>LASSO regression, GAM / sLM, Decision trees, logistic regression, 1 to 2 layer neural network, Random forest classifier, GAM and gradient boosted decision trees (LightGBM), SVM, linear regression, Multi-label gradient boosted tree, k-nearest neighbor and bagged decision tree</i>), Wizard of Oz
Explanation types	Local feature importance (<i>saliency maps, word highlighting, LIME, SHAP, sensitivity analysis MOEA/D...</i>), Rule-based, Example-based (<i>MMD-critic, nearest neighbours, manual inductive explanation...</i>), Counterfactual (<i>LORE, other...</i>), Textual (<i>in natural language: expert-generated or automatic</i>), Uncertainty estimation, Other Global (<i>distribution of values, decision tree, output visualisation</i>)
User expertise	Domain expert, Machine learning expert, Lay user, Researcher
Tasks and domains	Artificial task (<i>sentiment analysis of book and beer reviews, prediction of fat content in a food image, prediction of traffic accidents in a country...</i>), Law and regulation (<i>child welfare screening, identity recognition, recidivism prediction</i>), Business and finance (<i>credit scoring, house price estimate</i>), Education, Leisure (<i>chess, music recommendations</i>), Healthcare, Others (<i>application to lose weight, profession prediction, image recognition...</i>)

Table 3.1: Coding book used for the analysis of the corpus.

Identification of cognitive biases. To identify by name the cognitive effects that were discussed in the papers we reviewed, we either took the wording used in the papers, or relied on external taxonomies [Kahneman, 2011, Kahneman et al., 1982], surveys (e.g., [Kliegr et al., 2021]), and on our own knowledge of cognitive biases, specifically when the bias was not named explicitly. For a few cases we coined a phrase to be able to refer to the effect under study (e.g. “pre-use algorithmic optimism” [Springer and Whittaker, 2019]).

3.4 Results

This section presents the results of the analysis of the articles studied. First, we give an overview of the biases identified (RQ1). We then examine the stated mitigation strategies as well as the research methods used to identify them (RQ2, RQ3 and RQ4). For the sake of brevity, we do not systematically provide the definitions of the biases we examine, but the interested reader can refer to the lexicon provided in Appendix A1.

The first contribution of this work is to answer our RQ1 and identify the cognitive biases encountered in our corpus, along with the context in which they were found, namely the explainability technique that was used, the domain, the task, and the user type. We identified a list of cognitive biases in Appendix A1. The list presents all the expressions and concepts found in the corpus, but we recognise that some concepts may overlap and represent the same underlying cognitive mechanism.

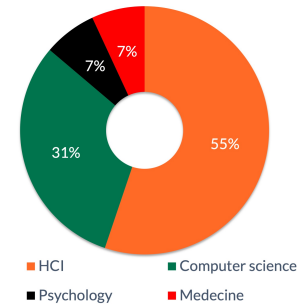


Figure 3.2: The distribution of the corpus across disciplines.

We then analyzed the way these biases were presented in the articles reviewed, revealing four main ways cognitive biases affect or are affected by the use of explainable AI systems for decision-making.

3.4.1 Overview

Figure 3.3 presents the different categories of explanation techniques that were seen in our corpus (in the middle). Each link represents a connection made in the literature between an explainability technique and a cognitive bias or between a cognitive bias and a mitigation technique. The legends in color underlined by arrows indicate how and in what direction the links should be read (e.g. "XAI techniques should adapt to explanatory heuristics"). The pale and wide links indicate that the bias or constraint applies more generally to all XAI methods. We identified more connections between biases and mitigation strategies but show only the most supported ones for brevity.

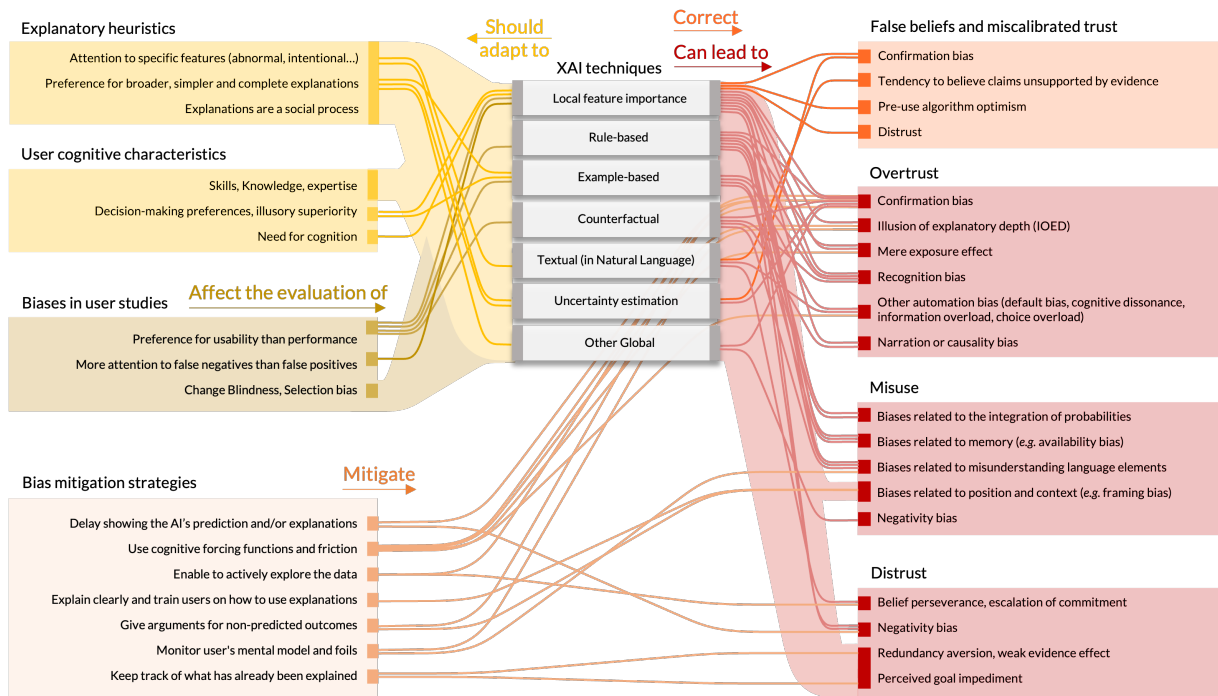


Figure 3.3: Summary of the cognitive constraints, biases and mitigation strategies discussed in the papers included in our corpus (n=38).

The first type are **heuristics and characteristics of users that should affect how explainable AI systems are designed**. They are listed in the **yellow boxes** in Figure 3.3 (top-left corner). They include all the explanatory heuristics that people use when explaining or receiving an explanation. These explanatory heuristics are well documented in psychological works on the human explanation process [Lombrozo, 2007, Miller, 2019]. Unlike the other types of cognitive biases discussed in our survey, these explanatory heuristics are not considered to lead to errors. On the contrary, they were simply presented as neither good nor bad but

merely cognitive architecture constraints to be taken into account before designing explainability techniques. We present them in Section 3.4.2.

The second type of cognitive biases are **those caused or exacerbated by explainability, and which can lead to erroneous decision-making**. They are presented in the **red boxes** on the right of the diagram in Figure 3.3. Among these, we find cognitive biases that lead either to overtrust, distrust, or to misusing the explanation. We present these in Sections 3.4.3, 3.4.4 and 3.4.5.

The third category are **cognitive biases that were successfully corrected by explainable AI**. They are presented in the **orange box** in Figure 3 (top-right corner). In Section 3.4.6, we review successful examples of using an explainability technique to address a false belief that was observed with non-explainable AI systems.

The fourth category we identified are **cognitive biases which can distort how XAI techniques are evaluated in user studies**. They are presented in the **brown box** in Figure 3.3 (middle left). Prompted by Doshi-Velez and Kim [2017], recent attention has been focused on approaches to evaluating explanations, with some researchers arguing for the need to test explanations with users [Poursabzi-Sangdeh et al., 2019], and others cautioning against doing so, concerned that cognitive biases could skew evaluations and mislead the XAI field [Herman, 2019]. We take stock of these cognitive biases in Section 3.4.7.

Finally, the bias mitigation strategies mentioned in the corpus are presented in the **pale orange box** (bottom-left of Figure 3.3). The identified biases leading to overtrust, distrust and misuse of explanations of AI systems are summarized in Table 3.4.3.

3.4.2 *Cognitive mechanisms explanations should adapt to*

Explanatory heuristics

In this section we summarize the cognitive (and biased) ways in which people select causes, evaluate, and ultimately trust explanations. As the term “bias” usually refers to errors in judgment and we do not consider such cognitive mechanisms as errors, we use the term “*explanatory heuristics*”. Unlike the cognitive bias of the other categories in Figure 3.3, in this class, the explanatory heuristics are inherent to the explanation process and help humans select some events as being relevant causes out of a potentially infinite causal chain of events [Hilton, 1988]. As presented in Section 2.3 in Chapter 2, explanatory heuristics were mainly examined by reviews such as [Miller et al., 2017], but also by primary studies focusing on explainability desiderata such as simplicity and completeness.

Attentional heuristics. People pay attention to some causes more than other to form explanations [Miller, 2019, Lombrozo, 2006, Malle, 2004].

Specifically, people tend to focus on causes that are abnormal, intentional, that point to the responsibility of individuals, that are necessary, sufficient and robust. Further, the studies in our corpus show that people select and assess causes according to confidence estimates [Bhatt et al., 2021, Miller, 2019, Wang et al., 2019a], demographic features [Liu, 2021] and inherent features [Bekele et al., 2018, Miller, 2019].

Bhatt et al. [2021] stress the importance of showing confidence estimates of AI prediction. They argue that people need to assess uncertainty to make decisions, relying on prospect theory [Kahneman and Tversky, 1979]. In social interactions, we are used to estimating the confidence level of a person's assertion based on their tone and other social cues. These cues are not applicable in human-AI interactions, hence the need to explicitly state AI's confidence levels. However, Bussone et al. [2015] nuanced that view by empirically demonstrating that "*the amount of system confidence had only a slight effect on trust and reliance*".

Liu et al. [2021] report on people's tendency to focus on demographic features such as race or age in feature-based explanations. We can hypothesize that may be due to the discriminatory potential of these features and may be linked to either the severity of the consequences of weighting in these variables or to the **availability bias**². This is consistent with earlier observations on trust in automation, that trust depends on the severity of the consequences of failure, *cf.* Section 3.2.

Another interesting example of incorporating these attentional biases into the design of XAI techniques is [Bekele et al., 2018], which used the inherence bias—a human tendency to focus on inherent features instead of extrinsic ones to explain a phenomenon—to select explanations for person re-identification systems.

² Human tendency to rely on information that comes readily to mind (such as information seen recently in the press) when evaluating a situation [Kahneman, 2011].

Preference for broad, simple, complete explanations. Additionally, existing work on explanation desiderata has evidenced that people look for specific qualities in explanations (*cf.* Section 2.3). In our corpus, we also observe such preferences for "broad" [Miller, 2019, Shimojo et al., 2020, Woodcock et al., 2021], "simple" [Abdul et al., 2020, Miller, 2019, Shimojo et al., 2020, Zytek et al., 2021] and "more complete" [Kulesza et al., 2013] explanations. However, the preference for simple and complete explanations raises several ambiguities. While it is unchallenged that simpler explanations are more comprehensible and readable [Abdul et al., 2020, Fürnkranz et al., 2020]—some researchers even show that interpretability is inversely related to explanation length [Fürnkranz et al., 2020]—they can also be received with skepticism by users [Bussone et al., 2015, Fürnkranz et al., 2020, Kulesza et al., 2015]. Similarly, Kulesza et al. [2013] argue that more comprehensive explanations help to significantly improve participants' mental models, but other work [Bussone et al., 2015, Szymanski et al., 2021] found complete explanations can lead to overreliance [Woodcock et al., 2021]. Shimojo et al. [2020], Woodcock et al. [2021] contend that coherent and broad explanations are preferred, with scope being even more important than simplicity, consistently with Lombrozo's point of view in cognitive science that broader and simpler explanations are better [Lombrozo, 2007]. Based on these findings, it can be challenging to gauge the right level of complexity in explanations.

Some suggested general principles such as not providing explanations that are too complex to be readable [Fürnkranz et al., 2020] or adjusting to the level of “completeness” to each user and context [Woodcock et al., 2021].

Social expectations. Furthermore, Weld and Bansal [2018] support Miller’s view [Miller, 2019] that explanation is a social process and state that adopting more “social” explanations would be highly beneficial to provide more relevant explanations. Through the process of dialogue, social explanations can be used to identify each user’s specific knowledge gap that needs to be explained.

Woodcock et al. [2021] highlight the impact of considering the explainee’s prior knowledge and the foil in her question that needs to be addressed. They show that explaining a disease to a user of an AI-powered chatbot who possesses prior knowledge of that disease has little impact on her trust. Then, **tailoring explanations** to addresses specific users’ questions has an important impact on trust. For that reason, some researchers have argued for more interactive explanations. However, there is some concern in the articles of our corpus that interactive explanations may lead to overtrust or overreliance [Liu et al., 2021].

Moreover, people tend to attribute human traits to machines, and therefore tend to expect that AI systems use the same communication framework as humans [Miller, 2019, Weld and Bansal, 2018]. This was already highlighted in early research on trust in automation [Lee and Moray, 1992, Glikson and Woolley, 2020]. In 1992, Lee and Moray explained that people tend to anthropomorphize machines and attach more importance to system characteristics than to system behaviour, as they would do when calibrating human-to-human trust. This is also known as the **correspondence bias**, whereby we tend to explain behaviour in terms of motives, traits and intentions, and underestimate the influence of external factors [Gawronski, 2004].

User individual characteristics

Some studies showed that certain individual characteristics of users impact the way explanations are received. Broniatowski [2021] stressed the importance of considering individual differences to design meaningful explanations.

Skills and expertise. The author considers the effect of skills such as numeracy— mathematical ability—, having a computer science background, or reading skills—which enable users to better “*extract the gist from narratives with poorly defined causal structures*” [Broniatowski, 2021]. The studies in our corpus also identified major differences in the way explanations are received depending on traditional classifications of user expertise. Experts have a greater ability to extract relevant information, follow efficient and trained reasoning paths, and generally avoid overreliance and overtrust [Broniatowski, 2021, Kahneman and Klein, 2009, Szymanski et al., 2021, Simkute et al., 2020]. Novices are more exposed to overreliance [Simkute et al., 2020].

Personality traits. Browniatowski also reviews the effect of certain personality traits on explanation reception. One aspect to consider is the Need for Cognition (NFC), which refers to an individual's desire for mental effort and can be quantified using the NFC scale [Broniatowski, 2021, Bućinca et al., 2021]. Additionally, Schaffer et al. [2019] discussed how **illusory superiority**³ makes people less likely to seek advice and may be linked to higher susceptibility to cognitive overload.

People also differ in the way they make decisions. Some tend to rely on their gut feeling, while others prefer to think long and hard. This trait can be measured through the Cognitive Reflection Test (CRT) [Broniatowski, 2021]. This echoes Coba et al. [2019]'s results. Coba et al. used a Choice-Based Methodology [Louviere et al., 2010] and eye-tracking measurements to reveal that people's various decision making styles impact how they perceive hotel ratings—shown as "collaborative explanations". People of the "maximizer" type were more prone to insensitivity to sample variance and choice overload.

³ Refers to psychological observations where low-skilled people felt a sense of superiority which made them less likely to rely on advice [Schaffer et al., 2019]. Also known as the Dunning-Kruger effect.

3.4.3 *When explainable AI leads to overtrust*

As studies in automation show, overtrust phenomena such as automation bias and complacency may arise with automated and AI systems. These mechanisms can be exacerbated by explainable AI, as studies in our corpus show. The interested reader can refer to the lexicon in the Table A.1 of the Appendix of this thesis for definitions of the cognitive mechanisms and biases in bold in the text.

According to the **mere exposure effect** [Kliegr et al., 2021], the sheer presence of an explanation increases confidence in the machine's prediction. This effect was evidenced in [Eiband et al., 2021, Fürnkranz et al., 2020, Lai and Tan, 2019], with lay users, rule-based and local feature importance explanations, by demonstrating that random or placebo explanations increase trust.

Several papers examined user's **bias for completeness** [Bussone et al., 2015, Fürnkranz et al., 2020, Kulesza et al., 2013, Lai and Tan, 2019, Szymanski et al., 2021]. For example, Fürnkranz et al. [2020] showed that users found longer explanations more plausible than shorter ones. This is consistent with [Bussone et al., 2015] which showed that giving a fuller explanation in the context of a medical diagnosis led to overreliance issues. Lai and Tan [2019] demonstrated that additional details including irrelevant ones improved user's trust in AI predictions. Szymanski et al. [2021] contended that the additional details contained in visual explanations compared to textual ones can increase users' misattributed trust. Finally, Szymanski et al. [2021] showed that lay users were more exposed to confirmation and completeness bias than machine learning experts when faced with visual explanations of a reading time prediction algorithm.

"Giving a fuller explanation of the facts used in making a diagnosis had a positive effect on trust but also led to overreliance issues, whereas less detailed explanations made participants question the system's reliability and led to self-reliance problems."

[Bussone et al., 2015]

These articles provide several avenues for addressing the bias for completeness problem, including by combining the use of textual and visual explanations [Szymanski et al., 2021] or by providing arguments against the machine's suggestion [Bussone et al., 2015].

Some mentioned the possibility that more complete explanations are more likely to contain elements that the user recognizes, thus contributing to the persuasive effect through the **recognition bias**⁴ [Kliegr et al., 2021].

⁴ Recognizing information makes the user more likely to trust the explanation [Kliegr et al., 2021].

In a healthcare application, Wang et al. [2019a] also reported that doctors who considered the AI prediction before making their own diagnosis fell into confirmation bias and relied on backward reasoning.

Another bias studied in the corpus is the phenomenon called "**illusion of explanatory depth**"⁵, coined by Koehler [1991] and evidenced in the explainability literature by Chromik et al. [2021] using local feature importance (SHAP [Lundberg and Lee, 2017]) explanations. They prompted users to self-explain so that they would realize that they knew less about the concept being explained than they had originally imagined. We can also perceive this effect in [Kaur et al., 2020, Naiseh et al., 2021b] which mentions "superficial" and "rush understanding".

⁵ People think they have a much deeper understanding of how complex concepts work than they actually do.

Several articles in our corpus emphasized that experts were particularly affected by **narration or causal bias**.⁶, including researchers who attribute causal meaning to saliency maps [Atrey et al., 2020], data scientists who make false narratives about how SHAP and GAM explanations work [Kaur et al., 2020] or domain experts in the domain of child welfare screening using counterfactuals [Zytek et al., 2021]. The authors mainly called for incorporating knowledge-based narratives in explanations. Atrey et al. [2020] encouraged researchers to use direct experimental evidence to back up their claims. In our corpus of articles, narration bias was linked to overreliance on explanations, following the same logic as confirmation bias and backward reasoning [Wang et al., 2019a]. People used narratives to make sense of the predictions of AI systems, which reinforced their trust in them. In [Zytek et al., 2021], counterfactual explanations lead users to mistake correlation for causation and develop flawed causal narratives.

⁶ Tendency to interpret information as being part of a larger story and to assume causal relations in the events of that story [Betsch et al., 2015].

Other biases related to complacency. Several studies reported tendencies from participants to over-rely on AI's predictions [Bansal et al., 2021, Bussone et al., 2015, Danry et al., 2023, Liu, 2021, Lai and Tan, 2019, Naiseh et al., 2021b].

Using an AI aid for chess, Bayer et al. [2021] demonstrated that chess players displayed a **default bias**⁷, that is, users tended to prefer the default option suggested by the AI. This behavior may overlap with the

⁷ Tendency to accept a presented default option (almost similar to status quo bias).

concept of automation bias demonstrated in early studies on automation. This suggests that offering AI predictions as a default option is probably a flawed strategy if we want users to actively critique and challenge AI decisions.

In a task called the Diner's Dilemma game, Schaffer et al. [2019] demonstrated automation bias towards AI recommendations. The authors did not find explanations to be an effective remedy.

Additionally, Danry et al. [2020] discussed the "**cognitive dissonance**" effect⁸—as study participants called it—and ties it to cognitive overload in a fake news detection task. When given a suggestion by the AI, the study participants were inclined to follow the AI's suggestions, even though they knew they might have opposing personal beliefs. Explanations reinforced that effect. In this study, AI suggestions were explained through arguments of why a claim is supported or not by evidence, in plain language and spoken to participants through an earpiece.

⁸ Having two opposing and coexisting beliefs, leading to cognitive conflict and psychological stress.

3.4.4 *When explainable AI leads to distrust*

Our corpus also contains articles discussing under reliance issues, which we refer to as "distrust". These were manifested through various aspects of overconfidence in one's abilities or choices, such as "**the escalation of commitment**"⁹ evidenced with chess players receiving text-based explanations [Bayer et al., 2021], the "**illusion of validity**"¹⁰ evidenced with domain experts [Simkute et al., 2020] or "**illusory superiority**" [Schaffer et al., 2019] for lay users with low levels of cognition.

⁹ The tendency to remain committed to a choice made, even though one understands with newer information that it leads to undesirable results.

Several works have highlighted the role of user expertise in distrust problems. Domain experts have developed cognitive routes that enable them to make quick and accurate decisions in environments that are "regular" enough to be predictable [Kahneman and Klein, 2009]. Their intuition is therefore more sophisticated than a lay user's "System 1" [Kahneman, 2011]. Simkute et al. [2020] highlight Klein [1988]'s findings that experts make decisions intuitively, with little uncertainty, and rarely consider more than one option. While useful heuristics, this reasoning also make experts more prone to **belief perseverance** [Koehler, 1991] or **algorithm aversion**¹¹, especially when faced with contradictions from the machine's predictions [Simkute et al., 2020]. In addition, user studies involving domain experts often focus on decision-making contexts that are high-stake, time-limited and stressful, as it is the case in the critical industries such as healthcare. This may explain the reluctance of experts to engage in explanations. Naiseh et al. [2021b] argue that experts in critical domains are in a serious state of mind, where they tend to perceive additional information as "**goal impediment**".

¹⁰ Tendency to overestimate one's ability to accurately interpret and predict results when analysing a data [Kahneman, 2011].

¹¹ "People erroneously avoid algorithms after seeing them err" [Dietvorst et al., 2015].

Negativity bias¹², was found to affect everyone including non-expert users. It can lead to significant trust loss when showing the weaknesses of the system early through explanations [Nourani et al., 2021, Kliegr et al., 2021, Shimojo et al., 2020, Zytek et al., 2021]. Nourani et al. [2021] suggest controlling what types of predictions users see when first interacting with the system.

¹² A tendency to pay more attention to negative features.

3.4.5 *When explainable AI is misused*

This section analyses other cognitive patterns present in AI-based decision-making. These patterns are not consistently correlated with overtrust or distrust, but instead display a misapplication or misunderstanding of the explanation. This leads to a poor calibration of trust.

Related to the integration of probabilities. In their review of biases related to rule-based explanations, Kliegr et al. [2021] described several cognitive biases related to people's difficulty to integrate probabilities such as **base rate neglect**¹³ or **conjunction fallacy**¹⁴ [Kliegr et al., 2021]. Fürnkranz et al. [2020] further evidenced that people (lay users in this case) tend to ignore the statistical significance of a statement, a phenomenon called **insensitivity to sample size**. Miller [2019] stressed that probabilities don't matter to people—a claim somewhat disputed by [Bhatt et al., 2021] if uncertainty estimates are probabilities—and that explanations should focus on causal relationships.

Related to memory. Wang et al. [2019a] discussed **representativeness**¹⁵ and **availability bias** in the context of medical diagnosis, and proposed showing prior probability and prototypes of outcomes to mitigate these.

Misunderstanding language elements. Biases leading to misusing the explanations can also be due to misunderstanding some elements of the language [Kliegr et al., 2021] that is commonly used in explanations such as the logical operator "AND" in rules [Fürnkranz et al., 2020], Boolean logic in counterfactuals [Zytek et al., 2021], or confidence scores when it is ambiguous what they refer to [Bhatt et al., 2021].

Related to position and context. Nourani et al. [2021] discuss the **primacy effect**¹⁶. They suggest controlling the type of predictions users observe when first interacting with the system [Nourani et al., 2021, Kliegr et al., 2021].

Additionally, Branley-Bell et al. [2020] explore user biases towards explainable AI system in a healthcare application. The research findings indicated that users exhibited greater trust in the system's accuracy when a malignant diagnosis was provided and explained, as opposed to when a benign diagnosis was given. Unlike the negative bias we examined in Section 3.4.4, this occurrence of **negative bias** leads to poor trust calibration rather than distrust. Here, trust is based on an irrelevant factor. Similarly, Mohseni et al. [2021a] observed that "*users pay less attention to false positive explanation errors and in turn, are more critical for false negative explanation errors*". This may be related to people's tendency to see false positives as less harmful than false negatives, and therefore relates to people's attention to the severity of the consequences of failure when calibrating trust in automation [Culley and Madhavan, 2013].

3.4.6 *When explainable AI corrects false beliefs*

Other explainability researchers have examined the extent to which explainable AI can successfully mitigate the cognitive biases that arise in

¹³ "The tendency to underweight evidence provided by base rates" [Kliegr et al., 2021].

¹⁴ Estimating the conjunction of two statements to be more probable than one of the two statements.

¹⁵ The similarity of objects or events makes people disregard the probability of an outcome [Kahneman, 2011].

¹⁶ A tendency to form an opinion based solely on the first piece of information received.

decision-making with AI systems. As Liao et al. [2020] indicate, “users also consider explanations of the AI’s decision as potential mitigation of their own decision biases”. The literature on explainability frequently discusses broad notions of transparency as a potential tool to mitigate aversion bias, see [Park et al., 2021] for example. However, we exclusively consider research that focuses on explainable AI and substantiates claims about explanation’s ability to mitigate bias.

[Wang et al., 2019a] observed that explainability users in healthcare fell into **confirmation bias**, whereby they would pay more attention to information confirming an existing hypothesis, instead of looking for evidence of alternative possibilities. To mitigate this effect, they implemented an explainable AI system in which input attributions (feature-based explanations) are shown before the class attribution (AI’s hypothesis). Furthermore, as [Bhatt et al., 2020], [Wang et al., 2019a] argue for showing AI’s certainty estimates to mitigate overtrust effects.

Springer and Whittaker [2019] evidenced how users had positive expectations of the transparent system before using it. To be able to refer to it later, we call this phenomenon “**pre-use algorithmic optimism**”. Springer and Whittaker conclude that showing explanations progressively, in this case local feature importance, was important to prevent users from overestimating the capabilities of the system. They suggest presenting explanations gradually or only when requested, to prevent users from losing trust when their expectations about the system are contradicted.

Danry et al. [2020] designed an explainable AI prototype that was successfully able to correct people’s **tendency to believe persuasive claims that are not supported by evidence**. For each claim on a socially divisive topic such as immigration or poverty, an explainable AI device classified the claim as supported by evidence or not and provided an explanation of that evidence, e.g. “a majority of Americans support a ban on flag-burning because a poll conducted by CNN in June 2006 found that 56% of Americans supported a flag desecration amendment.”. People were better able to distrust unsupported claims and trust supported claims, although this sometimes caused cognitive dissonance problems. Additionally, people trusted less evidence supported by anecdotal and expert evidence instead of study evidence.

Further, Zytek et al. [2021] demonstrated through a user study the usefulness of their “Case-Specific Details” interface for domain experts to screen child welfare cases. The interface displays the local contribution of the factors pre-selected by users, which proved useful in **correcting experts’ lack of trust** in the model, and highlighting differences between human and AI logic.

To prevent users from relying on how similar the current situation was to a previously seen case (representativeness bias), Wang et al. [2019a] also suggest to show prototypes of other cases, either sorted per a metric of similarity, or accompanied with a dissimilarity metric. However, Zytek et al. [2021] evidenced that case-based explanations of examples similar to the current situation enhance people’s tendency to make decisions based on similarity.

Other studies suggested mitigation strategies to overcome systematic errors with explainable AI systems, without testing them experimentally.

For example, [Wang et al., 2019a, Lai and Tan, 2019, Springer and Whittaker, 2019, Buçinca et al., 2020] suggest to delay showing the AI’s prediction and/or explanations to enable users to form their own hypotheses. [Naiseh et al., 2021b, Buçinca et al., 2020, Simkute et al., 2020] propose to use cognitive forcing functions and friction to favor users’ active cognitive engagement. [Simkute et al., 2020, Wang et al., 2019a] argue for enabling exploration of raw data, and [Naiseh et al., 2021b, Bussone et al., 2015, Kliegr et al., 2021] propose to educate users and clearly explain how to use explanations. Lastly, [Bansal et al., 2021, Bussone et al., 2015, Wang et al., 2019a] recommend to give arguments for non-predicted outcomes to favor the consideration of alternative possibilities than the one suggested by the AI.

3.4.7 *When explanations are misevaluated*

Users’ stated preferences are not indicative of performance. Buçinca et al. [2021] warned against using proxy tasks to evaluate explanations through user studies, *i.e.*, tasks that consist in subjectively rating the explanations. They noted that people’s subjective preferences for explanations were not indicative of the performance they would exhibit in making decisions with these explanations. Instead, researchers should use real tasks. This observation was also evidenced in our corpus with local feature importance, rule-based, example-based, and counterfactual explanations [Buçinca et al., 2021, Liu, 2021, Szymanski et al., 2021].

More attention to false negatives than false positives. Focusing on saliency maps for image recognition, Mohseni et al. [2021a] showed that people pay less attention to explanations of false positives than explanations of false negatives. They also showed that people rate differently techniques that differ only in appearance. To address these biases, they designed a human attention baseline to evaluate saliency explanations without having to resort to user studies.

Furthermore, Sokol and Flach [2020] called for caution about the phenomenon of **change blindness** in user studies, namely the “inability to notice all of the changes in a presented medium”, especially in an image. To address it, any change should be highlighted or made salient. Researchers should also be wary of selection bias when selecting participants for user studies through Amazon Mechanical Turk, usually more computer literate than the ‘normal’ population [Barbosa and Chen, 2019].

To circumvent the problems associated with user studies, Mohseni et al. [2021a] presented a promising evaluation methodology. Leveraging human annotators, they developed human attention masks which can be used to evaluate model saliency explanations for image and text domains.

Year	Title	Authors	Venue	EXPLANATIONS...			THE PAPER INFORMS ON...		
				...lead to overtrust	...lead to distrust	... are misused	explanation evaluator	correcting biases	Other cognitive trait
2020	COGAM: Me	Abdul et al.	CHI	Red					Yellow
2022	Visual Analyti	Andrienko et al.	IEEE CGA			Red			
2019	Exploratory ni	Atrey et al.	ICLR	Red		Red			
2021	Does the Wh	Bansal et al.	CHI	Red					
2021	The role of de	Bayer et al.	Jo. of Decision Systems	Red	Red				
2018	Implementing	Bekele et al.	IEEE CVPRW						Yellow
2021	Uncertainty a	Bhatt et al.	AIES					Orange	Yellow
2020	User trust and	Branley-Bell et al.	HCCI			Red			Yellow
2021	Psychologica	Broniatowski	NIST Report						Yellow
2020	Proxy Tasks	Buçinca et al.	IUI				Yellow		
2021	To Trust or to	Buçinca et al.	ACM HCI Jo.	Red					
2015	The Role of E	Bussone et al.	2015IEEE ICHI	Red	Red				
2021	I Think I Get	Chromik et al.	IUI	Red					
2019	Decision mak	Coba et al.	IUI			Red			Yellow
2020	Wearable Re	Danry et al.	ACM Ahs	Red				Orange	
2019	The Impact of	Eiband et al.	CHI	Red				Orange	
2020	On cognitive	Fümkrantz et al.	ACM Machine Language			Red			Yellow
2020	Interpreting In	Kaur et al.	CHI	Red					
2020	The Effect of	Kim and Song	CHI			Red		Orange	
2021	A review of p	Kliegr et al.	Artificial Intelligence	Red	Red	Red			Yellow
2013	Too much, to	Kulesza et al.	IEEE VL/HCC	Red	Red				Yellow
2019	On human pr	Lai and Tan	FAccT	Red					
2021	Understandin	Liu et al.	ACM HCI Jo.	Red					Yellow
2019	Explanation in	Miller	Jo. of AI						Yellow
2021	Quantitative E	Mohseni et al.	IUI			Red	Yellow		
2021	Explainable F	Naiseh et al.	Computer		Red	Red			
2021	Nudging thro	Naiseh et al.	BESC	Red					
2021	Anchoring Bi	Nourani et al.	IUI			Red			
2019	I can do bette	Schaffer et al.	IUI	Red					
2020	How Does Ex	Shimojo et al.	Frontiers in Psychology						Yellow
2020	Experts in the	Simkute et al.	ACM DIS	Red					
2020	Explainability	Sokol and Flach	FAccT				Yellow		
2019	Progressive c	Springer and Whitt	IUI					Orange	
2021	Visual, Textu	Szymanski et al.	IUI	Red					
2019	Designing Th	Wang et al.	CHI	Red		Red		Orange	
2019	The Challeng	Weld and Bansal	Com. ACM						Yellow
2021	The impact of	Woodcock et al.	Jo. of Medical Internet R						Yellow
2021	Sibyl: Unders	Zytek et al.	IEEE TVCG		Red	Red			

Figure 3.4: The 38 papers in the corpus and a rough indication of whether the paper reports on over- or distrust effects of explanations, on the misuse of explanations, or on other explanation-related phenomena.

Cognitive biases	Ex. of evidencing strategies	Ex. of mitigating strategies
Leading to <i>overtrust</i>		
Mere exposure effect, Completeness bias, recognition bias, Confirmation bias, Illusion of explanatory depth	Study the correlation between explanation length and perceived plausibility [Fürnkranz et al., 2020], Ask participants to rate their own understanding before and after self-explaining AI predictions [Chromik and Butz, 2021], Study the effect of placebo or random explanations [Eiband et al., 2021]	Give arguments for non-predicted outcomes [Bussone et al., 2015, Wang et al., 2019a, Weld and Bansal, 2018], Delay showing the AI's prediction and/or explanations [Buçinca et al., 2021, Lai and Tan, 2019, Springer and Whittaker, 2019, Wang et al., 2019a], Use cognitive forcing functions and friction [Buçinca et al., 2021, Naiseh et al., 2021a, Simkute et al., 2020], Include uncertainty estimates [Bhatt et al., 2020, Bussone et al., 2015, Wang et al., 2019a]
<i>Related to causality:</i> Narrative bias, Overgeneralization, Causation vs. correlation, attention to demographic features	Ask participants to describe explanations, analyze free text answers and verbalizations [Kaur et al., 2020]	Incorporate human expertise into explanations [Andrienko et al., 2022]
<i>Related to complacency and information overload:</i> Default bias, Cognitive Dissonance, Choice overload	Observe user's degree of agreement with the AI with vs. without explanations [Danry et al., 2020], Measure the user's cognitive load using the NASA Task Load Index (NASA-TLX) [Kaur et al., 2020, Springer and Whittaker, 2019], Eye-tracking measurements [Coba et al., 2019]	Do not use too many explainability types [Zytek et al., 2021], Use user-centric approaches [Naiseh et al., 2021b]
Leading to <i>distrust</i>		
Escalation of commitment, Illusion of validity, Negativity bias, Familiarity bias, Perceived goal impediment, Redundancy aversion, Weak evidence effect	Observe the relation between subjective confidence, subjective comprehension, and positive and negative AI outcomes [Nourani et al., 2021], Ask participants to think aloud while they make decisions [Wang et al., 2019a]	Enable to actively explore the data [Simkute et al., 2020, Wang et al., 2019a], Use gamification and personalization [Simkute et al., 2020], Keep track of what has already been explained [Miller, 2019, Naiseh et al., 2021a], Control the predictions users observe in the training phase [Nourani et al., 2021]
Leading to <i>misusing the explanation</i>		
<i>Related to the integration of probabilities:</i> Averaging bias, Base-rate neglect, Conjunction fallacy, Disjunction fallacy, Insensitivity to sample size, Unit bias	Measure the correlation between the user's confidence and supporting evidence [Coba et al., 2019, Fürnkranz et al., 2020]	Reminder of probability theory, Use frequencies instead of percentages, Show support as an absolute number [Kliegr et al., 2021]
<i>Related to memory:</i> Representativeness, Availability bias	Analyze reasoning process through free text questions and think-aloud protocols [Wang et al., 2019a, Zytek et al., 2021]	Show prior probabilities of outcome and examples of decision outcome [Wang et al., 2019a]
<i>Related to misunderstanding of language:</i> Misunderstanding of the inverse, of 'and', Boolean logic, confidence scores Analyze free text responses [Zytek et al., 2021]	Clarify the meaning of language elements to only one group of participants [Fürnkranz et al., 2020]	Clearly communicate what the presented information means [Bussone et al., 2015], State only true statements for the presentation of Boolean elements, including by negating false ones [Zytek et al., 2021]
<i>Related to timing and context:</i> Framing bias, Primacy effect, Anchoring bias	Measure the perceived reasonableness of explanations and the performance of users at a task under different explanation framing conditions [Kim and Song, 2020, Nourani et al., 2021]	Describe the uncertainty of both positive and negative outcomes [Bhatt et al., 2021], Control the kind of predictions users observe in the training phase [Nourani et al., 2021]

Table 3.2: Cognitive biases exacerbated by explainable AI and examples of evidencing and mitigating strategies.

3.4.8 Explanations tend to increase unwarranted trust

Overall, the studies in our corpus show a general tendency for explanations to increase trust, even when it is unwarranted, *i.e.* the AI is not trustworthy. For example, Bansal et al. [2021] note that “*explanations are interpreted as a general sign of competence*” and that “*explanations increased the chance that humans will accept the AI’s recommendation, regardless of its correctness.*” Nourani et al. [2021] also find that “*In all conditions, explanations increased confidence in the user’s estimations*”.

As illustrated in Figure 3.4, our corpus analysis revealed that explanations resulted in overtrust in 18 studies, while 6 studies reported distrust effects of explanations. Additionally, 12 studies identified cognitive biases that led to miscalibrated trust (it is not clear in which direction, overtrust or distrust). Although a broad range of cognitive biases have been discussed in the literature on explainability, it is possible that these biases may overlap and share common underlying trust mechanisms. For example, anchoring bias and confirmation bias may be two sides of the same coin when calibrating trust in explainable AI predictions. Central to the cognitive issue is the timing of when the explanation is presented to the user: whether it is before or after the user has formed her own opinion. Similarly, earlier investigations into trust in automation first distinguished between automation bias and complacency, eventually finding that these two phenomena largely overlap.

3.4.9 Important factors for appropriate trust: a Bayesian approach

Central to calibrating trust in explainable AI systems is how people reconcile AI predictions and their explanations with their prior knowledge [Chen et al., 2023, Shimojo et al., 2020]. This “belief reconciliation” process is related to the process of evaluating explanations according to coherence, or generality as described in [Miller, 2019]. The problem has also been framed in a more rational way in terms of probabilities as detailed in [Shimojo et al., 2020]. Shimojo et al. [2020] argue that “the [explainability] problem is one of updating posterior probability”. According to the authors, the Bayesian approach¹⁷ can be described as “the update of the probability that a cause induced an event after taking into consideration new information of the event.” In other words, the explainability problem in Bayesian terms consists in assessing the probability of an explanation to be true knowing an AI prediction ($P(C|E)$, the “posterior”), using the probability of the AI prediction to be true ($P(E)$, the “marginalization”), the probability of the cause presented in the explanation to be true ($P(C)$, the “prior”) and the probability of the AI prediction to be true given that the explanation is true ($P(E|C)$, the “likelihood”).

However, the authors also note that in practice, humans tend to disregard Bayes’ rule and estimate “subjective posterior probability” according to cognitive biases [Kahneman and Tversky, 1979].

The studies in our corpus reveal what appear to be recurrent and sig-

¹⁷ Bayes’ rule [Phillips and Edwards, 1966]:

$$P(C|E) = P(C) \frac{P(E|C)}{P(E)}$$

nificant trust factors involved in belief reconciliation, which ultimately lead to trust calibration. Furthermore, individuals' ability to reconcile prior knowledge and critically examine the coherence of explanations appear to be limited by three aspects: individuals' prior knowledge, the "probability" that a cause presented in an explanation is the cause of the AI prediction, and human and individual cognitive and attentional capacities. We examine the important trust factors in our corpus in terms of these three aspects of the belief reconciliation problem.

Prior knowledge. Several studies in our corpus highlighted the importance of **user expertise, task expertise and task familiarity** on the way people calibrate trust in explainable AI systems [Bayer et al., 2021, Bussonne et al., 2015, Zytek et al., 2021]. For example, Bayer et al. [2021] note that "*experts use explanations to resolve their disagreements. In contrast, novices lack expertise, which makes them reliant on the opinions of third parties, and rather than question these opinions, they tend to use them to learn (Gregor and Benbasat, 1999).*" Following the Bayesian approach, this expertise would enable users to assess the prior probability that a cause presented in an explanation is true ($P(C)$), or that an AI prediction is true ($P(E)$). For example, Bhatt et al. [2020] highlight the importance of showing **estimations of the AI's confidence**.

Explanation likelihood. Similarly, the quality and persuasiveness of the explanations providing information to update beliefs plays an important role for people to infer the posterior probability that a given explanation C is the cause of an AI prediction. Specifically, the papers we reviewed shed light on the importance of **explanation completeness** [Kulesza et al., 2013]. Out of 6 studies in our corpus that reported distrust effect, all were related to either user expertise (experts trusted less AI systems) or explanation's lack of completeness (incomplete explanations decreased trust in AI systems).

Cognitive and attention capacity. In addition, certain trust factors are linked to cognitive overload and limitations of human attention, *i.e.* the "capacity" and "attentional" problems described by Zerilli et al. [2019]. These factors include the **timing and framing of explanations, users' motivation and individual characteristics** such as need for cognition [Buçinca et al., 2021, Broniatowski, 2021] of decision-making preferences under choice overload [Coba et al., 2019]. As presented by Kim and Song [2020] and Nourani et al. [2021], timing and framing of explanation have an important part to play in human's ability to revise prior knowledge. These conditions seem to be decisive in activating confirmation or a narrative bias [Wang et al., 2019a, Bansal et al., 2021, Kim and Song, 2020]. Bansal et al. [2021] argue that "*by presenting an answer and accompanying justification upfront, and perhaps overlaid right onto the instance, our design makes it almost impossible for the human to reason independently, ignoring the AI's opinion while considering the task.*"

All these factors at play in the belief reconciliation problem may be related. For example, Schaffer et al. [2019] argued that lower cognitive ability as demonstrated by "illusory superiority" could be predicted by higher reported task familiarity.

Although our goal here is to identify high-level trust factors in explainable AI, we acknowledge that these may depend on specific AI applications and tasks. Liu [2021] note: *"Our work suggests that tasks may play an important role, and it can be challenging to understand the generalisability of results across tasks."* In addition, it was not always clear in the studies we reviewed what is the effect of explanations and what is the effect of AI predictions in users' trust calibration. For example, the illusory superiority bias leads to a general aversion to advice, and it is not clear whether explanations increase this bias compared to AI prediction alone [Schaffer et al., 2019].

3.5 Discussion

We present below a discussion of research directions we believe should be pursued in future work to address cognitive biases in XAI.

3.5.1 *Take into account cognitive mechanisms and biases in the design of explainable AI*

One of our aims in this work is to highlight the importance of considering human cognitive architecture in XAI design [Cummings, 2004]. This is common practice in the HCI field, but it may not have fully permeated a historically technical explainability field.

This may be a complex endeavour, however. Bayer et al. [2021] highlight the complexity of designing AI systems that takes into account opposing cognitive biases. On the one hand, they showed that users fell into a default bias when AI suggestions were presented at the same time as users were making decisions. On the other hand, participants fell into escalation of commitment when the AI suggestion came after they had made their choice. Kliegr et al. [2021] also mentioned the possibility that different cognitive biases could have opposing effects, such as information bias (leading to overreliance) and ambiguity aversion (leading to under reliance), thus emphasizing the need to consider biases in their context and to put them in relation to the user's knowledge. We also found contradictory results between Zytek et al. [2021], which found that example-based explanations for child welfare screening led to representativeness bias and Wang et al. [2019a], which presented prototypes of decision outcomes as a mitigation for the same bias. In addition, Lai and Tan [2019] warned about the "backfire effect" according to which *"corrections of misperceptions may enhance people's false beliefs"* [Nyhan and Reifler, 2010].

Lastly, there has been a surge of interest in interactive explanations recently, responding to the call to design explanations that fit the social process of explanation [Weld and Bansal, 2018]. However, concerns were expressed in [Liu et al., 2021] as interactive explanations were found to reinforce user's over reliance on AI suggestions. A possibility is that interactive explanations were more complex to interpret in Liu et al. [2021]'s study, leading to information overload.

Overall, more work is needed on the effects of interactive explanations, of bias mitigation measures and on identifying opposing biases and backfire phenomena.

3.5.2 *Clarify the normal vs. problematic biases with empirical and normative work*

Which cognitive biases need to be mitigated? In this review, we identified some cognitive biases as being neutral heuristics, *i.e.* "normal" ones, inherent to the process of explanations. Instead of mitigating those biases, some argue that they should be taken into account in the design of explanations [Miller, 2019, Weld and Bansal, 2018], for example by providing explanations as social processes or by adopting contrastive explanations. However, there is a blurred line between biases XAI needs to adapt to and those that need to be mitigated. It goes back to the important question posed by Weld and Bansal: "*Should an explanation system exploit human limitations or seek to protect us from them?*". Lakkaraju and Bastani [2020] argue that by exploiting certain human cognitive biases, such as preferences for relevant or familiar features, trust could be manipulated. Conversely, Miller [2019] explains that AI explanations should be contrastive, simple and when applicable delivered in the form of a dialogue, *i.e.* interactive. Clarifying which biases are normal and which are undesirable appears to be important for moving the XAI field forward. To that end, more empirical work on the benefits and drawbacks of incorporating cognitive constraints into explanation is needed.

Further, not only do we need more empirical research into user biases in explainable AI, but we also need more theoretical and normative work to distinguish genuinely biased cognitive processes from those that are normal. Such a distinction seems difficult to make without normative evaluations referring to the correctness of decisions and the inherent quality of the decision process for the users, including his or her level of participation. In fact, recent work has advocated for "functional" models of cognition, which differ from the "deficit" model of cognition such as the dual system theory [Kahneman, 2011]. These more contemporary models highlight that cognitive biases exist for good reasons, and often produce "good" rather than "bad" decisions, and study how heuristics help to make people better decision makers. Much of this research questions the conventional wisdom that intuition/heuristic thinking ("system 1 thinking") is "quick and dirty" while reasoning ("system 2 thinking") is slow and good. For example, Gigerenzer [2023]'s work shows that intuition is quick and error-prone, while reasoning is slow and just as error-prone. Normative work to help researchers and XAI designers decide whether, how and in which priority different biases need to be addressed should also keep in mind these more contemporary models of cognition.

3.5.3 *Detail taxonomies of user groups with cognitive factors*

Recent efforts to tailor explanations to the task at hand, the user's goals, knowledge [Coba et al., 2019, Szymanski et al., 2021, Woodcock et al., 2021] and specific needs [Simkute et al., 2020, Wang et al., 2019a], in order to meet the user's understanding, would be improved by taking into account the individual personality traits and specific skills we have mentioned in Section 3.4.2. Future work could consider how the current high-level groups of explainability users (currently categorized per AI expertise or role in AI system) could be detailed with this cognitive information, [Mohseni et al., 2021b, Suresh et al., 2021, Tomsett et al., 2018] highlighting cognitive biases each user group may be prone to. Broniatowski [2021] also suggested that the explainability field should strive to identify the individual factors that influence explainability in each user community.

3.5.4 *Improve our perception of users' reactions to XAI*

Several authors have advocated that we need a better perception of social and emotional behavior of users to be able to correct errors in their reasoning and their mental models of the system [Akata et al., 2020, Chromik et al., 2021, Woodcock et al., 2021]. As a first step towards this, we highlighted some methods to evidence biases in Table 3.4.3. Notably, what seems to be a good practice for controlling for the mere exposure effect is using placebic explanations or randomly generated explanations as a baseline [Eiband et al., 2021, Nourani et al., 2021]. Then, cognitive load can be measured through the means of the TLX workload assessment method [Kaur et al., 2020, Springer and Whittaker, 2019], eye-tracking measurements [Coba et al., 2019] or through the number of cognitive chunks and a subjective measure encompassing the reading time, the self-reported load and memory performance (how well the user remembers the explanation) [Abdul et al., 2020]. In addition, we frequently encountered the use of qualitative analyses in our review, such as think-aloud protocols [Naiseh et al., 2021b, Springer and Whittaker, 2019, Szymanski et al., 2021, Wang et al., 2019a], useful as pre-studies but not generalizable (they involved from 12 to 20 participants in our corpus), or the analysis of free text comments, which can be implemented more easily on a larger scale [Naiseh et al., 2021a, Szymanski et al., 2021, Zytek et al., 2021]. Further, the ability of XAI systems to capture users' mental states could be complemented by a memory of these states and a memory of what has already been explained [Miller, 2019, Naiseh et al., 2021b].

3.5.5 *Focus on strategies beyond XAI: contextualization, training, timing, cognitive forcing...*

Various work in our corpus mentioned the need to pay more attention to other interaction design choices [Buçinca et al., 2021, Zhang et al., 2021] beyond the choice of an explanation method. These include contextual information, training, timing, framing, and other specific strategies to mitigate cognitive biases. For example, Simkute et al. [2020] suggested

the use of gamification strategies in low-stakes environments to address the lack of motivation of some users, and the use of feedback and controls in high-stakes environments. Others stressed the need to clarify specific elements in the explanations. Bussone et al. [2015] proposed presenting how the explanations were derived, which Dazeley et al. [2021] calls "meta-explanations". Buçinca et al. [2021], Lai and Tan [2019], Wang et al. [2019a] suggested to delay showing the AI's prediction and/or explanations to decrease overreliance issues. Nourani et al. [2021] recommended to control the type of predictions that users observe when learning to use the system, during the initial instructions and training phase. Finally, Buçinca et al. [2021], Naiseh et al. [2021b] proposed cognitive forcing functions and friction-based strategies to address users' lack of curiosity. Cognitive forcing functions consisted in making users wait for the explanations, updating them or asking for them. The friction function designed by Naiseh et al. [2021a] consisted in asking the user to confirm that they did not want to review the explanation. All these strategies proved to be useful in decreasing user's unjustified trust, though it decreased their satisfaction in the system.

3.5.6 Give arguments against the prediction

The idea of explaining not only the AI's prediction but also alternative possibilities appeared in several papers [Bussone et al., 2015, Wang et al., 2019a, Weld and Bansal, 2018] as a way to counter automation bias. Wang et al. [2019a] recommended to support "*premortem of decision outcomes*", a reasoning consisting in trying to disprove a hypothesis. Bussone et al. [2015] highlighted comments from participants saying they wanted to see both positive and negative evidence for the suggested medical diagnosis. Finally, Bansal et al. [2021] envisioned an AI that would play "*a devil's advocate role, explaining its doubts, even when it agrees with the human*". They proposed a prototype of such an explanation and found that while it was effective in informing the human that the AI might be wrong, it was not sufficient to reduce significantly errors related to overreliance. One of the main challenges is getting users to come up with their own solution when they are informed that the AI may be wrong. Additional work is still needed to find the right kind of interaction that could help users detect that the AI is wrong [Bansal et al., 2021], but the direction seems promising, notably for two reasons. First, it reminds us of the adversarial structure of a judicial system where two parties (a defense attorney and a prosecutor) present opposing arguments. Implementing such "adversarial explanations" could increase societal trust in the AI-aided decision process. Second, a necessary condition for free will is the availability of alternative possibilities, or the ability to "choose otherwise" [McKenna and Coates, 2021]. Therefore, showing alternative explanations to the decision-maker helps with sustaining her autonomy and accountability.

3.6 *Limitations*

Since our goal was to provide insight into how the XAI field has considered cognitive biases to date, we used a systematic search methodology. This allowed us to cover a broad sample of articles on XAI. However, it is possible that some articles did not use our general search terms on cognitive biases and focused on specific types of cognitive biases in XAI. Our paper augmentation is limited by potential biases in the authors' view of the XAI field. To continue this line of research on cognitive biases, future review work could focus on specific biases, such as "automation bias". Evidently, our list of cognitive biases cannot be considered as the finite list of biases affecting explainable AI systems, there are numerous others in the cognitive science literature which may be worth studying in the context of XAI. Moreover, it was quite difficult to assess the generalizability of the results presented in our corpus. To address this limitation, we tried to preserve the context in which these results were obtained — explainability technique, user type, and task type. However, it is possible that these results depend on more granular details. Finally, we leave it for future work to produce more interactive versions of a heuristic map such as the one we present, in a similar fashion as Suresh et al. [2021]. This could facilitate the tracking of cognitive biases that have been highlighted in the explainability literature and the contexts in which they have been highlighted.

3.7 *Conclusion*

In this chapter, we presented a scoping review of 38 papers — from a corpus of 285 papers — to investigate what kind of cognitive biases were identified in the presence of explainable AI systems. In addition, we conducted a qualitative analysis of these papers, providing a map of the different cognitive biases and revealing the context in which they occur, specifically with which XAI technique, type of user, and AI-assisted task.

Furthermore, our mapping shows the different ways in which these biases affect XAI-assisted decisions. We highlighted the ways in which explainable AI can often lead to overtrust, or distrust, the latter occurring either with expert users or with incomplete explanations. Explainable AI has sometimes been misused by end users, who have been shown to misunderstand some linguistic elements or probabilities, to rely on irrelevant information from their prior experiences, or to be sensitive to the framing and timing of the explanation. Cognitive biases can also affect the way explanations are evaluated in user studies. However, explanations can still contribute to correct cognitive biases such as confirmation bias, correcting overly positive expectations of AI systems or believing persuasive claims that are unsupported by evidence.

Overall, explanations tend to have a positive effect on trust. This can lead to an "**explanation paradox**", where explanations may increase users' unwarranted trust and make them more vulnerable, rather than empowering them with information about the AI's prediction. Important

factors in calibrating trust in explainable AI systems include user expertise, task expertise and task familiarity, estimation of the AI's confidence, explanation completeness, timing of explanations and users' motivation and individual cognitive characteristics (need for cognition, rational or intuitive decision-making style...). We provided several directions for future work that pave the way for meeting users' cognitive needs.

In the next chapter, we explore whether interactive explanations can effectively address this search for human-centric and even 'human-like' explanations.

Chapter 4

Towards "human-like" explanations: the promise of interactivity

"Explanations should be interactive, allowing the explainee to revise and consolidate some previous background knowledge."

Confalonieri et al. [2021]

TO ADDRESS the trust calibration challenges posed by cognitive biases, we have stressed the importance of the human-centric approach, and to take into account the human cognitive explanation process. Both empirical research surveyed in Chapter 3 and theories in psychology and sociology support this view. Interactivity in explanations has been advanced by recent work on human-centered XAI as a promising way to align with cognitive human architecture and support reconciliation with prior beliefs [Chen et al., 2023, Wang et al., 2019a, Adadi and Berrada, 2018, Miller, 2019, Langer et al., 2021, Arya et al., 2019, Longo et al., 2020, Atakishiyev et al., 2020, Confalonieri et al., 2021, Krause et al., 2016]. However, empirical research on interactive explanations is still emerging, and it is still unclear whether they really live up to their promise. In particular, it remains uncertain whether they are able to correct the overtrust and overreliance effects that "normal" explanations tend to produce, as seen in Chapter 2, or whether, on the contrary, they exacerbate them.

In this chapter, we examine what are the different types of interactive explanations and to what extent they align to the human explanation process through a detailed scoping review. We also take stock of their effect on user trust and reliance on AI systems and other user-based metrics. Section 4.1 outlines the motivation for the survey presented in this chapter and research questions. Section 4.2 describes the relevant related work and Section 4.5 lays down the survey methodology used. The results are a taxonomy of the interaction types for explainability, and an analysis of interactive explanations' usage, evaluations and effects. They are presented in Section 4.4. Finally, Section 4.5 discusses open challenges

for interactive XAI.

4.1 Motivation and research Questions

Building on natural sciences theories is common practice in HCI. The objective is to design artefacts that align with human cognitive processes. Recent work in HCI has focused on aligning explanation design with people's cognitive explanation process, resulting in the advocacy of more interactive explanations [Longo et al., 2020, Atakishiyev et al., 2020, Confalonieri et al., 2021, Arya et al., 2019, Krause et al., 2016]. Relevant results in the social sciences for explanations are summarized succinctly in Figure 4.1.

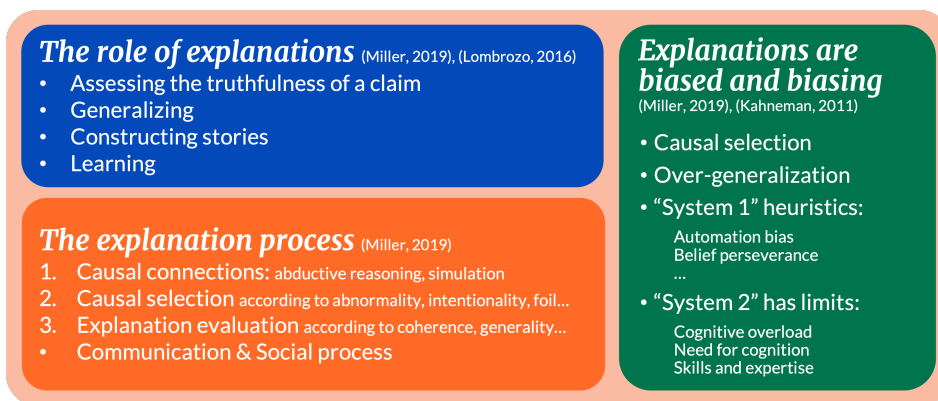


Figure 4.1: Summary of the role of explanations, the process by which we construct and present explanations and the biases involved in explanations.

For example, people expect explanations to be provided in a personalized request-response pattern [Graaf and Malle, 2017]. In addition, as seen in Section 2.3 presenting explainability literature in the social sciences, one does not ask "why P?" but rather "why P *and not* Q?" [Hesslow, 1988, Lipton, 1990, Millicamp et al., 2019]. That is to say, explanations are contrastive. These explanation characteristics call for ways to enable user interaction with explanations, and to make explanations more responsive. Rohlfing et al. [2021] emphasises that these considerations are still largely unaddressed in the explainability literature and calls for a 'social practice' of explanation in which explainers and explainee co-construct understanding.

Furthermore, research in the field of education shows that interactivity plays a fundamental role in learning [Sims, 1997, Barker, 1994]. Barker [1994] describe interactivity as "a necessary and fundamental mechanism for knowledge acquisition". Although the objectives of an explainable AI user may not include long-term learning, they generally revolve around acquiring knowledge about the AI system. We can therefore consider the problem of explainability as a learning one, reinforcing our assumptions about the important role of interactivity.

The term "interactive", however, can refer to many different kinds of user interactions. According to Miller [2019], the ideal interaction model

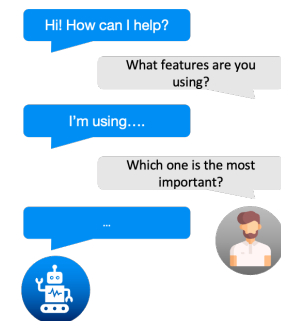


Figure 4.2: Illustrative example of interactive explanation: "Conversational XAI" enables users to interact with users through natural language.

follows a human-like dialogue structure, where the AI agent is able to answer a series of questions. Other types of user interaction have been implemented by XAI researchers, such as simulating the black box with new inputs [Cheng et al., 2019, Chromik et al., 2021, Morrison et al., 2018], re-configuring the explanation space [Hohman et al., 2019], changing explanations [Khurana et al., 2021, Spinner et al., 2020], etc. However, these studies in XAI do not use a common vocabulary to designate different interaction types, making it difficult to study and draw general conclusions on the different forms of interactive XAI. The visualization (Infovis) [Yi et al., 2007, Keim, 2002, Roth and Mattis, 1990, Wilkinson, 2005, Amar et al., 2005] and other Human-Computer Interaction (HCI) [Sims, 1997, Rhodes and Azbell, 1985] communities have done extensive work on the classification of different modes of interaction. The explainability field is less mature. We believe that the explainability field would benefit from using a more precise and shared vocabulary to designate the different types of interactivity, taking inspiration from other HCI sub-fields.

Due to the increasingly large number of articles on XAI, researchers may be overlooking best practices and opportunities for interaction. To illustrate the complexity of designing interactions, Sims [1997] referred to it as "an art" requiring multiple considerations and a vast array of skills on the part of designers. This work aims at helping XAI system builders by centralizing examples of interactive explanations taken from various contexts (user expertise, XAI method, domain...).

Over the past few years, a growing body of work has been testing interactive XAI systems with real users, generating sometimes seemingly contradictory observations. Cheng et al. [2019] find that the possibility to simulate new predictions by changing input features improved user understanding compared to static explanations. However, concerns were expressed in Liu et al. [2021] because interactive explanations were found to reinforce users' overreliance on AI suggestions. One possibility is that interactive explanations were more complex to interpret in [Liu et al., 2021]'s study, leading to information overload. Another possibility is that understanding a model may not help much when the model and the user disagree. In short, explanations may not be so useful at helping people determine whether to trust one's own intuition or to trust the model output. At this stage, review work is needed to summarise the effects of interactive XAI from a user perspective, paving the way for subsequent systematic reviews to formally disentangle these findings.

In this work, we conduct a detailed scoping review on interactive and user-evaluated explainability systems. We survey two popular digital libraries for the HCI community: IEEE Xplore and ACM Digital Library. We are guided by four research questions.

RQ1: *What are the interactivity approaches that have been implemented so far in the explainability field?*

RQ2: *In what context, with what content, and in what form were the interactive explanations presented to users?*

RQ3: *What are the metrics used in user-based evaluations of interactive explanations?*

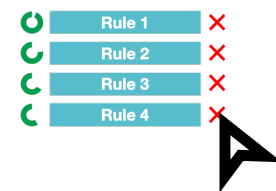


Figure 4.3: Illustrative example of interactive, rule-based explanation where users can create and modify rules.

RQ4: *What are the effects of interactive explanations on users' perception of explanations?*

To the best of our knowledge, we present the first review of the effects of interactive explainable AI on user experience.

4.2 Background

Below, we highlight work in HCI, XAI, and education that is relevant for our work. We also highlight, through these different strands of literature, reasons to believe that interactivity in explainability could help users in building sense and knowledge about models.

4.2.1 Interactivity in HCI

Defining interactivity proves challenging, and multiple definitions have been offered over time. Early work on interactivity defined it simply as the extent to which a user can "activate" [Sims, 1997] or "exert an influence" [Sundar et al., 2010, Steuer, 1992] on the technology being used, its form and its content. In 1997, Sims [1997] mentioned that "there appears to be no consensus of what interactivity actually represents or involves". Dix and Ellis [1998] and Foley et al. [1996] broadly define it using the keywords "communication between user and system" and "human-computer dialogue" [Yi et al., 2007]. In Infovis, Yi et al. [2007] view interaction techniques as "the features that provide users with the ability to directly or indirectly manipulate and interpret representations". The authors noted that Infovis systems were designed to communicate information from the computer to the user, but less so for the user to enter data, thus overlooking an entire aspect of interaction in HCI. Therefore, differences arise between HCI subdomains on how interactivity is defined. At first glance, it seems that the vision adopted by the Infovis domain could correspond to interactivity in XAI. In the explainability field as well, the user needs to manipulate, interpret and discover information about the model from explanations or raw data. In Section 4.4.1, we will examine how adapted the Infovis' view of interaction is to the XAI domain. Despite the lack of a consensual definition, Janlert and Stolterman [2017] state that "there seems to be a common sense understanding of interactivity as something fairly simple" that HCI researchers see as "the control and action between a human and an artifact or system."

However, defining the different types of interactions quickly complicates the task. Some studies have addressed it by proposing taxonomies of user-system interactions. Early ones attempted to provide holistic views of the interaction space in HCI; they focused on interaction *levels*, with the idea that "the higher the interaction level, the better the product" [Sims, 1997]. For example, Rhodes and Azbell [Rhodes and Azbell, 1985] introduced a three-level scale of interactivity, ranging from *reactive* to *proactive* to *coactive*. Schwier and Misanchuk [Schwier and Misanchuk, 1993] added two other dimensions to this taxonomy: functions (confirmation, pacing, navigation, inquiry, elaboration) and trans-

actions (keyboard, touch screen, mouse, voice). Sims' taxonomy [Sims, 1997] extends the two previous ones by intertwining functions and levels. It is presented as a scale from basic to complex with the following levels of interactivity: *object, linear, hierarchical, support, update, construct, reflective, simulation, hyperlinked, nonimmersive contextual and immersive virtual*. In the Infovis domain, there is typically no hierarchy between interaction types; however, taxonomies with finer granularity have been designed. For example, Yi et al. [2007] observes a difference of approach between system-centric taxonomies (including categories like "interactive linking and brushing" [Keim, 2002] or "navigating", *e.g.* zooming, panning [Wilkinson, 2005]) and user-task-centric taxonomies (including categories like "compare within relations" [Roth and Mattis, 1990] or "retrieve value" [Amar et al., 2005]). The taxonomy in [Yi et al., 2007] proposes to "connect user objectives with the interaction techniques that help accomplish them." It includes seven categories: *select, explore, reconfigure, encode, abstract/elaborate, filter, connect*. Yi et al.'s taxonomy has been extensively used and referred to in Infovis in the last decade.

4.2.2 *Interactivity in Explainability*

The call for more interactive explanations in XAI finds roots in results from the social sciences about how people communicate explanations and in the growing number of studies focusing on human needs rather than solely technical aspects. For example, Miller [2019] finds that "an explanation is an interaction between two roles: explainer and explainee". As such explanations should be thought as a social process, *i.e.* a conversation. The paper also mentions the rules that govern this interaction such as Grice's maxims [Grice, 1975] of quality (say only what is true), quantity (say no more than you need to), relation (say what is relevant to the conversation) and manner (say it in a nice way). Although it is easier to imagine these exchanges taking place in natural language, Tim Miller argues that this interaction can use other media such as images, keywords, or logical rules, while still respecting Grice's maxims. This work envisions what "human-like" explanations may look like, noting that users of XAI systems will expect explanations to be delivered in this manner.

The line of research on interactive XAI has begun to investigate how to tailor explanations to users. Work pertaining to the technical aspects of XAI also identifies the importance of such "user-centric" explanations. [Sokol and Flach, 2020, Schneider and Handali, 2019]. Numerous papers have emphasized the need for explanations that are tailored to the context, audience and purpose of the explanation [Doshi-Velez and Kim, 2017, Adadi and Berrada, 2018, Ras et al., 2018, Ferreira and Monteiro, 2020, Došilović et al., 2018]. Schneider and Handali [2019] reviewed XAI studies focusing on personalization. For each paper in their corpus, they documented personalized explanation properties (complexity, content and presentation), personalization granularity (to each user or per category of user) and personalization automation (manual or automatic). Additionally, they observed that personalization of explanations can be

either iterative or one-off, with user information being collected once prior to showing explanations [Schneider and Handali, 2019, Sokol and Flach, 2020]. While the personalization of explanations is particularly important given the role of explanations in filling one's specific knowledge gaps, we believe there is a greater granularity of interaction to explore beyond the categories mentioned in [Schneider and Handali, 2019].

As seen in Section 2.4, more and more HCI researchers have been investigating user's needs for XAI using standard HCI methods [Kou and Gui, 2020, Lim and Dey, 2009, Penney et al., 2018, Sun et al., 2022]. These efforts have resulted in numerous examples of sophisticated interactive interfaces integrating sometimes complex XAI techniques. For example, the strand of research called "conversational XAI" made significant strides in providing explanations in natural language to a wide range of user questions [Sokol and Flach, 2020, Hopenstal et al., 2021, Hernandez-Bocanegra and Ziegler, 2021].

4.2.3 *Interactivity for learning and sensemaking*

Explainability is also deeply connected to results in educational research. The parallel seems natural, as the field of explainability aims to improve human understanding of algorithms, or for machines to teach humans about their breakthroughs [Schneider and Handali, 2019]. According to Roussou [2004], many educational researchers agree that interactivity plays an important role in learning, notably by supporting "learning by doing". Amthor [1992] argues that "*people retain about 20% of what they hear; 40% of what they see and hear; and 75% of what they see, hear, and do*". This follows the constructivist approach, which emphasizes the need for people to build knowledge by testing and simulating new situations that have meaning for them [Dewey, 1903, Roussou, 2004]. Kent et al. [2016] demonstrates through quantitative user studies "*the role of interactivity as a process of knowledge construction*" and further asserts that interactivity patterns inform on the actual learning process of an individual. Evans and Gibbons [2007] find that interactivity promotes deep learning by stimulating users' cognitive engagement in the learning process. To tie more concretely these results to the explainability field, we can draw a parallel between the processes of learning, knowledge construction and that, closely related, of sensemaking. Cabrera et al. [2022] studied the cognitive process of sensemaking of models, and highlighted that "*understanding of models is an iterative and ongoing process*", motivating the need for their XAI system to be interactive. In this case, the sensemaking—or knowledge construction—, comes from the ability to iterate between the discovery of instances, the formation of hypotheses, their evaluation, etc.

4.3 Methodology

To review the role of interactivity in XAI, we conducted a scoping review drawn from an initial extraction of 716 papers, narrowed down to our final corpus comprising 48 articles. In this section we detail the characteristics and different phases of the survey method.

4.3.1 Review type

This chapter presents a scoping review [Arksey and O'Malley, 2005], as presented in Section 3.3.1. The scoping review methodology corresponded to our objectives of identifying, mapping, reporting and discussing the available evidence on interactivity in XAI. As in the previous chapter, we also rely on a standardized search and selection methods from the systematic review methodologies [Page et al., 2021], as suggested in [Arksey and O'Malley, 2005] for scoping reviews, to ensure the replicability and transparency of our findings. We followed the paper identification, screening, eligibility evaluation and analysis procedure stages outlined in the PRISMA methodology [Page et al., 2021] to guarantee the quality of our search and selection process [Tricco et al., 2018].

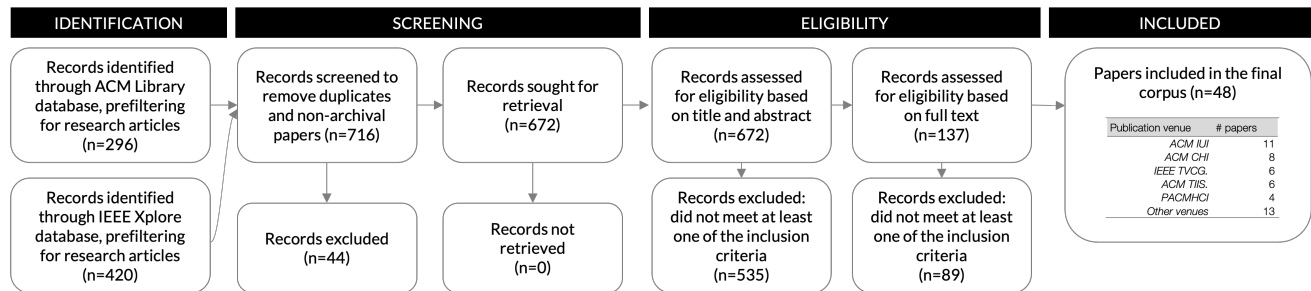


Figure 4.4: PRISMA flow diagram adapted from Page et al. [2021] giving an overview of the PRISMA 2020 survey guidelines, used for the search and selection phases of our scoping review.

However, our work goes beyond what is traditionally expected of a scoping review in particular in Sections 4.4.5 and 4.4.4, where we advance a summary of the effects of interactivity through Figure 4.19. We argue that this step enables us to better delimit gaps in the literature, and provide qualitative grounds for a following systematic review on a more restricted set of studies. This analysis is made possible through a minimal quality control of the included studies that we enforced through the exclusion of entries that were not published in a peer-reviewed conference proceeding or journal. However, a more thorough quality assessment of studies—which entails a restriction on the scope of the survey—should be performed in order to extract quantitative evidence about the effects of interactivity. Here, we aim at identifying the different types of results in the interactive explainability field and orientate further research. Section 4.6 discusses the limitation of the methodology in further detail.

For all these reasons, we refer to our type of review as a detailed scoping review.

4.3.2 *Corpus creation*

Identification. We focused on the ACM Digital Library and IEEE Xplore, two popular databases for the HCI community, which encompass prominent publishing venues for the explainability field (ACM CHI, ACM IUI, IEEE VIS, IEEE TVCG...). Consequently, we focused on XAI work that mainly—though not exclusively—pertain to the HCI community, rather than the computer science side of XAI. The main reason for this is that our focus was on interactivity and user studies—two topics finding roots in HCI. Moreover, the CS side of XAI has been historically and predominantly occupied with technical advances in XAI [Doshi-Velez and Kim, 2017], and has only very recently taken into consideration the user’s perspective. While we acknowledge that more interactive XAI systems have been emerging from the CS community recently, such as [Slack et al., 2022], interaction design has been quite distant from theoretical domains in computer science, as mentioned in [Abdul et al., 2018]. This led us to focus on HCI databases and leave out works published in purely AI conferences, such as NeurIPS, AACL, or CVPR, among others.

Our aim was to review different types of interactive explanations, focusing on how they are perceived by end users. Therefore, we narrowed our focus to work presenting an XAI interface and including a user-based evaluation of the XAI system. Note that there also exist non user-based evaluations of XAI methods. Doshi-Velez and Kim [2017] distinguish three evaluation strategies: application-grounded—testing explanations in real-world settings with domain experts—, human-grounded—testing explanations with lay users—, and functionality-grounded—testing explanations using metrics that do not require human feedback. The scope of our survey is limited to empirical studies with human subjects, as we are interested on the users’ perception of XAI systems. Providing insight into how people interact with XAI can guide practitioners in making more effective technical and design choices.

The keyword search was contextualized focusing on three dimensions: *AI Systems*, *Explainability* and *User studies*. The term "interaction" is ubiquitous in HCI¹, and as such we did not restrict our keyword search to this dimension, choosing instead to select articles on interactive explanations in the eligibility phase. Since we wanted to focus on articles whose main topic was AI, we searched for keywords representing AI systems and explainability dimensions in the Title, Abstract and Author Keywords fields. For the user study dimension, we searched the full text of the articles: we noticed that often, authors do not explicitly mention that they conducted a user-based evaluation in their abstract. The search results were limited to relatively recent articles (2015 or later), as XAI is a recent field of study, found to be expanding around 2016-2017 [Barredo Arrieta et al., 2020, Adadi and Berrada, 2018]. In addition, user-based evaluations and interest from the HCI community in the domain are even more recent [Doshi-Velez and Kim, 2017]. Using 2015 as a starting point, we are sure to capture the uptake in number of contributions in XAI.

In addition, we used ACM DL and IEEE Xplore filtering tools to narrow our search to research articles only. In ACM DL, we used the fol-

¹for example the CSS concepts section in ACM papers often include the term

lowing filter: All Publications/Proceedings/Content type/Research article AND All Publications/Journals/Content type/Research article, therefore excluding surveys, tutorials, introductions, editorials, newsletters, books, magazines, reports, encyclopedias, short papers, extended abstracts, posters, and other non-archival content. In IEEE Xplore, we used the filters Conferences and Journals, leaving out early access articles, magazines, books and standards. This step allowed us to make a first sorting of the non-archived articles, and facilitate the following phase of manual screening. For each record, the article title, authors, publication venue, and publication year were exported to an Excel spreadsheet. Below is the search query used (the wildcards * denote where we have retrieved the plurals and term variants):

AI systems => Abstract: (AI, artificial intelligence, machine learning, algorithm*) AND

Explainability => Abstract: (explainab*, explanation*, intelligib*, interpretab*, transparen*, XAI) AND

User studies => Abstract: (participant*, human-subject*, human evaluation*, human experiment*, user-stud*) AND

Date => 2015 or after AND

Journal or conference article => Non-archival records pre-filtered out.

Screening. One author deleted 44 records that were either duplicates or non-archival records that remained after the database filtering (primarily workshop entries and student consortia). This step resulted in a total corpus of 672 unique papers.

Eligibility evaluation. The remaining records were randomly assigned to three of the authors, who performed a two-phase eligibility assessment: a first one based on the title and abstract and a second, more in-depth one based on the full text. The first phase was primarily concerned with excluding recordings that were not focused on XAI (IC₁, IC₂), that did not include a human-AI interaction (IC₃), or that were a secondary study (IC₇). The second phase consisted of verifying IC₄, IC₅, and IC₆, since full-text viewing was required to assess these criteria. The inclusion criteria were the following:

IC₁ *XAI focus.* The paper's contribution is in the explainability field;

IC₂ *XAI system.* The paper shows an implementation of an XAI systems;

IC₃ *Human-AI interaction.* The paper is in the field of human-AI interaction (works in human-robot interactions are excluded);

IC₄ *User-based evaluation.* The paper presents an evaluation of its explainability approach using human-grounded evaluation [Doshi-Velez and Kim, 2017];

IC₅ *Human-computer interface.* The paper describes the interface that was presented to the human users evaluating the XAI system;

IC6 Interactivity. The explainability approach presented in the paper is interactive, meaning the user can interact *with the explanation* (requiring another interaction than that with the interface to perform a specific task)²;

IC7 Primary study. The paper is not a review nor a position paper.

After the three reviewing authors had completed the eligibility phase, an external reviewer was asked to apply the above criteria to a subset of 67 articles randomly selected from the base of 672 papers, representing 10% of the papers. Inter-rater reliability was 92%, and the remaining disagreements involved mostly cases in which the external reviewer included the articles when the authors did not. However, we believe that the extra step of reviewing the full text in detail is what justified the exclusion of the items that the external reviewer included.

One of the articles included in our corpus [Gu et al., 2021] was an analysis of an external primary study that did not match our keywords because it did not mention explainability-related terms in the abstract, but it met our inclusion criteria. We therefore replaced the secondary study with the primary study [Yan et al., 2020].

Eventually, 48 papers met the inclusion criteria and were included in the final corpus.

4.3.3 Analysis and coding book

Analysis process. The synthesis methodology we used in this review is an emerging synthesis [Schick-Makaroff et al., 2016], more specifically a narrative account of included studies, as is usually the case in scoping reviews [Arksey and O'Malley, 2005]. To support this analysis, we use a concept matrix and a charting approach to provide basic numerical summaries of the extent, nature and distribution of the studies included in the review.

Following Webster and Watson [2002], we created a concept matrix for the analysis of the interactivity landscape in the explainability field. The matrix is organized into four dimensions, whether the concepts relate to the *context* of the explanation, its *content*, its *communication*, or its user-based *evaluation*. Three authors independently coded and classified the articles included in the final corpus. For the dimensions context and content, the categories used for coding were predefined. In the communication dimension, only the concept of "representation" had a set of predefined categories. With respect to the type of interactivity, the different categories were intentionally not preset in advance and each of the three coders created their own categories after encountering an interactive explanation implementation. We did this because our goal was to create new categories that matched the range of interactivity types provided by the corpus. The authors then reviewed the resulting categories and discussed how to reconcile them into a taxonomy of interactivity types adapted from well-known existing ones [Yi et al., 2007, Sims, 1997]. A similar approach was taken for the evaluation portion of the matrix. As new types of evaluations were found, new categories were created.

² Some examples of papers excluded because of IC6 are [Bansal et al., 2021, Buçinca et al., 2020, Dominguez et al., 2019], which present static explanations to end-users, although the user interface to perform a downstream task may be interactive.

We grouped together concepts that were very similar (such as *explanation utility* and *explanation usefulness*). Finally, evaluations that were used only once in the corpus were regrouped in the "other" category of the matrix. The authors of this work discussed and shared the definition of the notions during several meetings. One author reviewed all the papers and corresponding codings to check the consistency of the two other reviewers' coding with their own, and subsequently consolidated the matrix. Below we detail the different concepts we have analyzed in each dimension.

Dimension	Code	Reference
Context		
Domain	Law and Civic, Healthcare, Business and Finance, Education, Leisure, Artificial, Generic, Other.	[Lai et al., 2021]
Audience	Domain experts, AI experts/Data scientists, Non-expert, Other.	[Lai et al., 2021]
Data type	Image, Video, Audio, Tabular, Natural language, Sequential data.	NA
Content		
XAI focus	Raw Data, Output, Model Limitations, Model Confidence How?, Why?, Why not?, How to?, What if?, What's the difference with?, Context.	[Lim and Dey, 2009, Liao et al., 2020, Sun et al., 2022]
XAI method	Local Feature Contribution, Decision Rules, Sensitivity Analysis and Partial Dependence Plot, Example-based, Saliency mask, Concept-based, Surrogate model, Counterfactual, Wizard of Oz.	[Lai et al., 2021]
Communication		
Interactivity	Clarify, Arrange, Filter/focus, Reconfigure, Simulate, Compare, Progress, Answer, Ask.	[Yi et al., 2007, Sims, 1997]
Representation	Chart, Table, Text, Rules, Directly on the data structure, Other.	NA
Evaluation		
Comparison	No explanation, Static explanation, Other, No baseline.	NA
Evaluation measure	Perceived usability, Perceived usefulness, Understanding, Perceived explanation length/quantity, Time spent interacting with XAI system, Trust, Cognitive load, Performance at task, Learning, Predicted accuracy, Perceived control, Perceived fairness, Perceived transparency, User skepticism, Other. <i>Only for evaluations using static or no explanation as a baseline: Higher than, Same as, Lower than [the baseline], Other.</i>	NA

Context. We retrieved the environment in which the explanations for each item were designed: domain, audience, and data type. The domain and audience categories are adapted from those found by [Lai et al., 2021] in their survey of AI-assisted decision making tasks. This allows us to see if the interactive explanations are well distributed across these contextual concepts.

Content. To analyze the content of the explanation, we searched for the explanation *focus*, which described the type of information that was provided to the user, and the explainability method used to extract it. The list of explanation focus points was adapted from Lim and Dey [2009], Liao et al. [2020] and Sun et al. [2022]'s classifications of user questions in XAI. The categories of the explainability method were adapted from [Lai et al., 2021].

Table 4.1: Codebook used to retrieve information from the corpus with four dimensions: [explanation] *context*, *content*, *communication* and *evaluation*, their corresponding sub-dimension and reference from which codes were inspired from.

Communication. Communication refers to the form in which the explanation was provided to the user, including the type of interaction used and the type of visual representation of the explanation. The categories of interactivity are described in more detail in Section 4.4.1. The categories of representation were kept general as they were not the focus of this study.

Evaluation. One of the main challenges in XAI is how to measure the quality of an explanation [Colquitt and Rodell, 2015]. User-based methods have been an increasingly adopted approach following calls such as Doshi-Velez's [Doshi-Velez and Kim, 2017] to take user perspective into account instead of just technical constraints. While "human-grounded" evaluations may have drawbacks such as sampling bias or change blindness [Sokol and Flach, 2020], they do inform how end users understand, perceive, and use explanations. This approach also has the advantage that standard questionnaires are shared by researchers to measure concepts such as trust (using the McKnight framework), satisfaction, understanding, cognitive load (using NASA-TLX), etc. We also retrieved the baselines (no evaluation, static evaluation, other explanation, etc.) used to evaluate the presented explanation in each empirical study. This makes it possible to compare the results of multiple studies and to get an overview of assessments of interactive explanations. For each evaluation in the corpus that used either static or no explanation as a baseline, we reported the results according to four categories: higher than, same as, lower than the baseline, or "other", which referred to more nuanced results dependent on other external factors, or to evaluations that did not rely on a defined baseline.

4.4 Results

4.4.1 Interactivity types in explainability: Select, Mutate, Dialogue with

Let us now describe the categories of interactivity in XAI that we have identified in our corpus. We took inspiration from other existing taxonomies of interactivity [Sims, 1997, Yi et al., 2007] to define these categories. This section addresses our RQ₁ and RQ₂.

Nine different categories of interactivity in XAI emerged from our analysis. Following Yi et al. [2007] and Roth and Mattis [1990], we formulated the categories so that they express interaction actions that correspond to user intents. We adapted some categories from Sims [1997] and Yi et al. [2007]. However, contrarily to Yi et al.'s taxonomy, the object of the interaction are explanations instead of datapoints. Explanations are larger constructs encompassing a visual representation, an input data range, an AI model's configuration (dataset, model type and parameters) and an explainability technique.

In addition to the categorisation of interaction types, we organized the taxonomy into three different groups corresponding to the type of support they provide for the human cognitive process of explaining.

This higher-level categorization is based on Miller's review of social science findings on properties of human explanations. Miller points out that explanations are selective, contrastive, and social. First, explanations are selective as they involve only a few causes in a large chain of causal events. Only a few causes address the explainee's question and are thus relevant. Then, explanations are contrastive as they are thought in contrast to a specific foil. People's questions are almost always "why" questions implying a foil: "why did P happened *and not* Q?" To assess the plausibility of a factor as a cause of an event, people then need to perform mental mutations, *i.e.* to cancel a factor which might have led to P and see if Q happens, or to consider situations where Q happened instead of P. This mental process is called the *mutability* of events and allows the formation of contrastive explanations. Finally, explanations are social because they are best understood in a conversation. The structure of the dialogue allows people to get specific answers to their "why" questions and corresponding foils, to ask follow-up questions and progressively fill the gaps in their knowledge.

Our proposed interactivity groups reflect the degree to which the interactive features enable these explanatory properties—selective, mutable, social. The three categories are: **select** (interactive features facilitate the selection of causes and the formulation of hypotheses), **mutate** (interactive features allow users to compare or simulate different configurations of the AI's inputs, outputs or parameters), and **dialogue with** (interactivity allows users to engage in a conversation with the XAI system). The resulting interactivity taxonomy is outlined in Table 4.2.

Below we describe in detail the nine different categories of interactive explanations, as well as three levels of interaction into which they fall.

Function	Category	Definition
Select	<i>Clarify</i>	Give additional information/explanations on demand
	<i>Arrange</i>	Choose and organize the explanation type(s), parameters and visual representation(s).
	<i>Filter/focus</i>	Filter the explanation according to an input/input metric.
Mutate	<i>Reconfigure</i>	Change the dataset, the AI model, AI model parameters and show me the corresponding prediction and explanations.
	<i>Simulate</i>	Change the inputs, the output or the dataset distribution and show me the corresponding prediction and explanations.
	<i>Compare</i>	Show me explanations of related or selected data inputs or outputs.
Dialogue with	<i>Progress</i>	Guide user through an explanation sequence.
	<i>Answer</i>	Give feedback, edit explanation components.
	<i>Ask</i>	Ask iterative questions and receive answers following a dialogue structure.

Select

The user may be able to select³ the information they wish to see by clicking on hyperlinks to display explanations on demand, by configuring the explanation space, or by filtering the explanation conditionally on an input metric. These interactions can help users formulate hypotheses and actively search for factors that may lead to causal explanations. As such, they enable explanations to be "selective".

Clarify. This subset of interaction capabilities enables the user to make on demand information appear, whether by clicking on or by brushing explanation components. In this approach, the user actively seeks answers to their questions, controlling what explanation to display and when it should be displayed. This set of interaction techniques is close to Yi et al. [2007]'s "elaborate" category. The analysis of our corpus revealed three main ways for a user to get clarification on something. First, users can navigate through a menu so as to choose the themes they want to know more about. Sims [1997] refers to this interaction technique as "hierarchical interactivity". Anik and Bunt [2021] is an example of this interactivity type. Second, explanations can be displayed after a user clicks on a link, following Sims [1997]'s "hyperlinked interactivity". One example is Sovrano and Sovrano and Vitali [2021]'s explanation system in which the user can click on a concept to get more information about it. With each click, a new window with an explanation appears, itself providing other hyperlinks about the notions used in the explanation. Finally, tooltips are convenient interaction techniques to provide clarifi-

Table 4.2: Two-level taxonomy of interactivity techniques in XAI, including a first level reflecting the type of support interaction techniques provide to the cognitive process of explaining, a second task-oriented level, and corresponding definitions.

³ A parallel can be drawn here with the "select" category from Yi et al. for the Infovis domain, which is defined as "marking something as interesting". Assuming we view this level of interaction as "marking an explanation as interesting", we found, however, several sub-categories of interaction types that could be used to support this. This justifies why we refer to it as a whole interaction level instead of just one category.

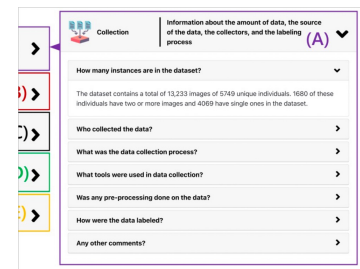


Figure 4.5: Example of the clarify interaction taken from [Anik and Bunt, 2021].

cations and additional details on a visualisation in a non-overwhelming way [Jin et al., 2020, Shi et al., 2019, Ahmad et al., 2019, Sevastjanova et al., 2021].

Clarify interactions also allow the explanation interface to be less overwhelming at first glance by disclosing explanations progressively. In a study on the progressive disclosure of explanations, Springer and Whittaker [Springer and Whittaker, 2019] note that "because transparency is provided 'on demand' this removes confusions and inefficiencies arising from spurious, unwanted explanations, and adjusts explanations to the users' requirements." They also observe that this on demand disclosure approach is able to adapt to the different reactions and expectations of each individual user.

Arrange. Arrange interaction techniques provide the user with the ability to organize the explanation space as desired by hiding or collapsing explanations and selecting the type of explanation to be displayed [Kwon et al., 2019]. It is similar to the "rearrange" category in Yi et al. [2007]. Instead of interacting for more information, (which corresponds to the *Clarify* category), here the user's goal is to configure the explanation space following their preferences. For example, in Liu et al. [2021], users can increase or decrease the number of highlighted words in the saliency-based explanation. In Collaris and van Wijk [2020], the user can chose the surrogate model used in the explanation along with the other parameters for that model.

Filter/focus.

Inspired by Yi et al.'s "filter" category, the *Filter/focus* class regroups controls that let the user zoom either on specific inputs of the AI model or subgroups in the the training or testing dataset. The user can therefore focus their attention on the explanation built from a restricted input space. The explanation interface presented in [Jacobs et al., 2021] is an example of a *Filter/focus* interaction technique where users (doctors) can filter explanations based on the presence of a specific symptom. In [Cheng et al., 2019], users can create and delete subgroups in the model's input data to see the corresponding explanations for each subgroup. VBridge [Cheng et al., 2022] and ExplainExplore [Collaris and van Wijk, 2020] provide the ability for users to select a subset of features to be used in an explanation. We also put in the *Filter/focus* class sorting functions, such as the one in Gamut [Hohman et al., 2019] which lets the user sort input features according to several feature metrics.

Mutate

Interactive explanations can allow the user to "mutate" causes, *i.e.* to test their hypotheses by simulating or comparing different situations. The resulting explanations are cumulatively selective and contrastive.

Reconfigure. This category includes a set of interactions that offer the possibility to modify the parameters of the AI model such as the dataset, the model type or the model parameters in order to observe changes on

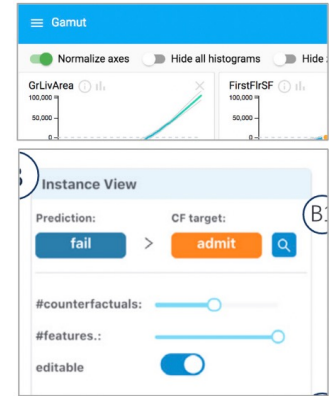


Figure 4.6: Examples of the arrange interaction taken from [Hohman et al., 2019] (top) and [Cheng et al., 2021] (bottom).

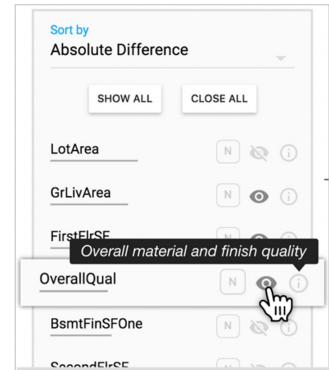


Figure 4.7: Examples of the filter/focus interaction taken from [Hohman et al., 2019] (top) and [Ming et al., 2019] (bottom).

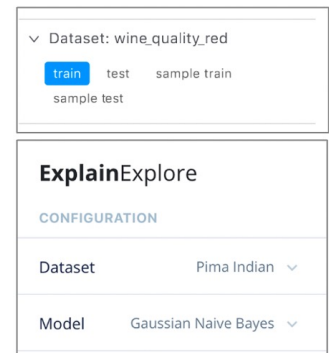


Figure 4.8: Examples of the reconfigure interaction taken from [Ming et al., 2019] (top) and [Collaris and van Wijk, 2020] (bottom).

the explanation. Users may want to evaluate the impact of these factors on the model's prediction and corresponding explanation to make sense of how the model works. This is especially true when explainability is used to assess the fairness of the model such as in [Yan et al., 2020] or [Lee et al., 2019]. The Silva explanation interface [Yan et al., 2020], similarly to IBM's AIF360 tool [Bellamy et al., 2019], allows the user to modify dataset attributes and sensitive inputs to see how it affects specified fairness measures. Various explanation components, including causal graphs and measures of feature importance, change based on the user's chosen dataset settings.

Simulate (inputs). Interactive explanations can be useful for users to test how changes in inputs affects local explanations and the outcome of the model. Understanding of a model then comes not only from static information about the AI algorithm, but also from the learning experience provided by repeated simulations of the model. Interactions in the *Simulate* category refer to mutations of the inputs of the AI model. Many articles in our corpus (18/47) have integrated this interactive feature, reflecting an appreciation of the XAI community for "learning by doing" [Roussou, 2004]. The simulation functionality is usually activated by sliders or drop-down lists and gives the user a local understanding of the model's behavior. Examples can be found in [Liu et al., 2021, Morrison et al., 2018, Ahn, Yongsu et al., 2022, Sevastjanova et al., 2021].

Compare.

This category gathers interaction techniques that are used to compare either (1) explanations for different inputs or group of inputs or (2) explanations for different predictions.

In the first case, the user can select the inputs or input groups to compare so as to analyze differences in the explanation. Connections, similarities and differences between the selected inputs or outcomes can be highlighted in the comparative explanations. Compare interaction methods would often use parallel coordinates graphs to ease the comparison between explanations. Hohman et al. [2019] give an example of an explanation view in which the user can see local explanations for two inputs they selected for analysis. The second case occurs when the AI model predicts several possible outcomes with varying levels of confidence. The user then usually wants to compare the explanations for each of the probable outcomes to assess their likelihood. Dodge et al. [2022]'s and Jin et al. [2020]'s systems are examples of this type of outcome comparison. In [Dodge et al., 2022], the user can tap on a game board (representing a game situation) to see its corresponding chance of winning and how it compares to the chance of winning from other game boards. In CarePre [Jin et al., 2020], doctors are users, and can explore in detail the records of a patient, as well as compare it with similar patients; their focus is on sequences of "events" (a patient enters the medical facility, a scan is performed, etc.). This allows the user to detect similar paths, and adapt treatment accordingly. This interaction class is inspired from Yi et al.'s "connect" category.

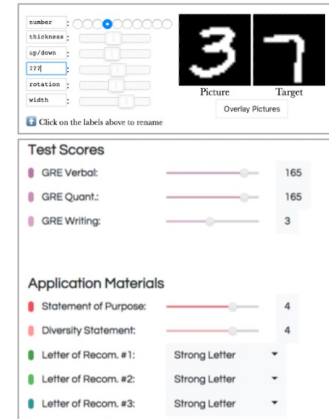


Figure 4.9: Examples of the simulate interaction taken from [Ross et al., 2021] (top) and [Cheng et al., 2019] (bottom).

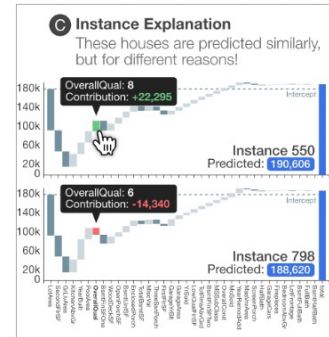


Figure 4.10: Example of the compare interaction taken from [Hohman et al., 2019].

Dialogue with

Interactivity can support the user in engaging in a dialogue-like structure. Information about the AI model is then given progressively and/or iteratively. The user could ask the system a question or give it feedback. These "dialogic" explanations are in line with the properties expressed by Miller for human-like explanations. However, there may be different degrees in which explanations are truly social, depending on the range of questions a system can actually answer.

Progress. The *Progress* interaction style is inspired by Sim's "linear interactivity" through which "the user is able to move forward or backward in a pre-determined sequence of instruction materials". The explanation is designed in several steps, and the user can click "next" or "previous" to navigate through the explanation displays. It is generally progressive, with basic information provided in the first few pages and more in-depth information presented in subsequent sections. This style of interactivity is reactive [Sims, 1997] and does not provide specific feedback to the user but instead lets them walk through the explanation at their own pace. The user can only control when the explanation is provided.

The *Progress* interaction style can be seen as the lowest level of "dialogic" explanations. It does not enable the user to ask nor answer questions but it follows some of the rules of a conversation [Grice, 1975] by providing sparse information progressively (maxim of quantity), and by predefining user questions that need to appear in the explanation guide (maxim of quality). The "next" and "previous" commands can be considered as the users' options to punctuate the conversation (compared to saying "ok tell me more" or "wait, what did you say").

Answer. While information flow in interactive XAI systems goes primarily from the machine to the user, like in Infovis systems [Yi et al., 2007], it can also be reversed, with users providing the system with feedback, corrections or information about the state of their mental models. These interactions can serve to increase users cognitive engagement (and activate their "System 2" [Kahneman and Tversky, 1979]) by challenging users. For example, in [Melsión et al., 2021], users (in this case children) are asked to click on the part of the image that they think had the most impact on the algorithm's prediction. This interaction type can also serve to improve the AI system by building on human feedback. Examples are [Jia et al., 2022, Shi et al., 2019] in which users are asked to improve the semantic meaning of the concepts learned by the algorithms, [Guo et al., 2022, Cheng et al., 2021] in which users can create or edit explanations—such as adding a new rule or correcting one, [Virgolin et al., 2021] in which users can indicate to the system their personal preferences about model interpretability, or [Hepenstal et al., 2021, Ghazimatin et al., 2021, Ghai et al., 2021, Spinner et al., 2020].

Ask. In [Miller, 2019], the ultimate level of interaction is a conversation where the user can ask the AI system anything they want. We can therefore view the *Ask* interactivity as the higher end of the interactivity scale for XAI. The conversational XAI research line has made some progress

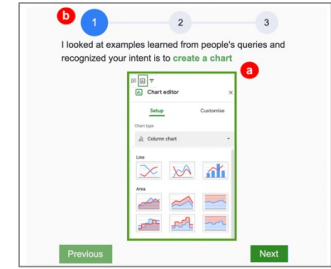


Figure 4.11: Example of the progress interaction taken from [Melsión et al., 2021].

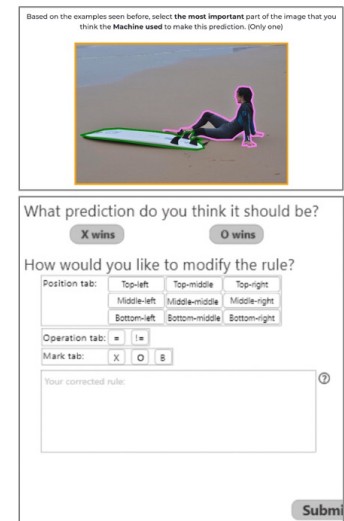


Figure 4.12: Examples of the answer interaction taken from [Melsión et al., 2021] (top) and [Guo et al., 2022] (bottom).

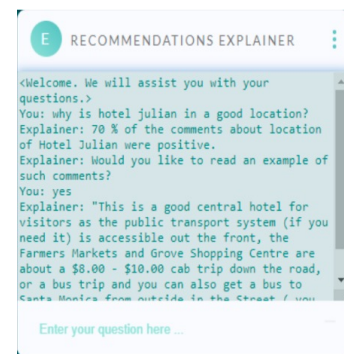


Figure 4.13: Example of the ask interaction taken from [Melsión et al., 2021].

in achieving such interactivity. For instance, [Hernandez-Bocanegra and Ziegler, 2021, Hepenstal et al., 2021] present logical dialogue maps to deliver explanations that answer users' questions. The challenge is to cover as wide a range of questions as possible. Note that this "dialogic" interaction between user and machine does not necessarily have to take place through natural language. As Miller stated [Miller, 2019], we could imagine an XAI system that answers the user's questions with images or other communication means. An illustration of this can be found in [Khurana et al., 2021], where the user submits a query such as "create a graph showing the predicted trend" and the XAI system responds with the desired graph.

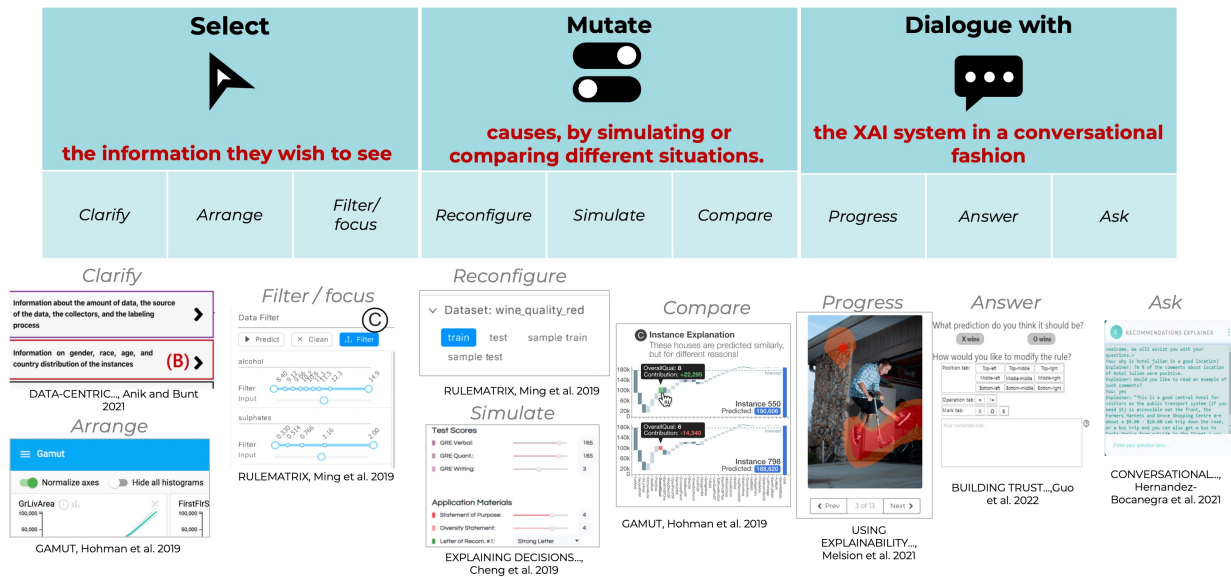


Figure 4.14: "Interactive XAI helps users. . ." Illustration of the taxonomy of interaction in explainability with screenshots from the corpus.

4.4.2 Context, content and form of interactive explanations

This section presents a qualitative analysis based on our conceptual matrix to address our RQ2.

Context. The work in our corpus is well distributed across the different domain categories constituted by [Lai et al., 2021] (cf. Figures 4.16 and 4.17). Notably, the corpus reflects a large number of studies (32/48 papers) implemented in real-world applications rather than in artificial or generic domains. Healthcare stands out as one of the most studied domains in the corpus.

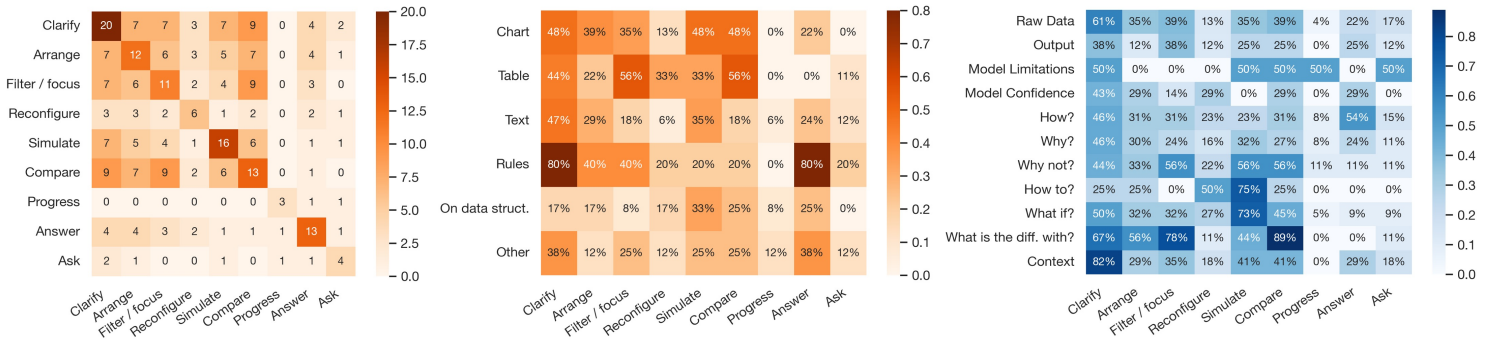
Some work [Bove et al., 2022, Cheng et al., 2019] expressed concern that too few studies focused on making explanations understandable to novices and that most current XAI techniques were only comprehensible to AI-educated users. Cheng et al. [2019] also argues that the majority of studies providing explanations to novices have been conducted in the context of generic tasks [Lai et al., 2021], *i.e.* computer science problems,

and are therefore not generalizable to real-world applications. In contrast to the first concern, we found that the majority of articles included in the corpus (27/48) were aimed at a general audience of non-expert users. This at least reflects an awareness of the field to design explanations with this user group in mind. In addition, 15/27 of these studies are in real-world application areas, including areas that may be considered sensitive—4 in legal and civil, 2 in healthcare, and 3 in business and finance. However, it is possible that the empirical studies included in our corpus targeted non-expert users for practical reasons, such as to solicit platform workers like those on Amazon MTurk [Guo et al., 2022, Hernandez-Bocanegra and Ziegler, 2021, Ross et al., 2021, Ghai et al., 2021, Cheng et al., 2019, Schaffer et al., 2015, Ribeiro et al., 2016, Wilkinson et al., 2021]. Nevertheless, some of these studies are primarily aimed at making the XAI systems more transparent and more accessible to a non-expert audience [Tsai et al., 2021, Springer and Whittaker, 2019, Yan et al., 2020, Szymanski et al., 2021, Anik and Bunt, 2021].

Regarding the data type used in our corpus, tabular and text data are predominant (79% of the studied papers). This points to an opportunity for the explainability field to empirically study interactive explanations using audio (only one paper discussed audio data [Anik and Bunt, 2021]), images, and video data.

Content. The interactive explanations in the corpus focused heavily on the "why?" user question recurring 37 times, and which can be answered by local feature explanations, the most commonly used explanation method in the corpus (26/48). We can see in Figure 4.15 (Right) how some interaction techniques were favored for specific types of user question. For example, quite logically, explanations addressing "what is the difference with?" were implemented with *Compare*, but also frequently with *Filter/focus* interactions. Context and raw data can be elaborated through *Clarify* interaction. "How to?" and "What if?" were facilitated through *Simulate* interactions. Model limitations were rarely presented in the studies (only twice). But perhaps a bigger opportunity for interactive explanations is the small numbers of papers addressing "how to?" questions. One example is [Ross et al., 2021] in which the user can change input "concept features" to see the adjusted output in real time and better understand the meaning of each "concept feature". However, we found only two studies enabling direct interventions on the model output [Jin et al., 2020, Dodge et al., 2022]. Such interventions (which would fall in the *Simulate* category cf. Table 4.2) could help the user characterize what kinds of contexts and situations are emblematic of a particular outcome, thereby addressing "how to?" questions. In addition, concept-based explanations, which are considered promising in the field for their human comprehensibility, were rarely used in the corpus [Kim et al., 2018, Koh et al., 2020].

Communication. The most used interaction techniques were *Clarify* and *Simulate*. These were frequently combined with *compare*, *Filter/focus* and *Arrange* as illustrated in Figure 4.15 (Left). The techniques *Progress*



and *Ask* were used in only three and four studies respectively, illustrating a trend in the field of interactive XAI towards complex, Infovis-type XAI interfaces rather than simpler step-by-step or dialog box interfaces. The matrix in Figure 4.15 (Left) shows this clear cut between the "Select" and "Mutate" interaction groups on the one hand, and the "Dialogue with" group on the other. The interactions techniques in the first two groups are frequently combined with each other, while the interaction styles in the latter group are less frequently used. In addition, these more "social" interactions were rarely combined with other interactions from the "Mutate" or "Select" levels. In particular, *Progress* was never used in combination with other "Mutate" or "Select" interaction categories. It would be interesting for future research to explore combining these as a way to take advantage of the social nature of "progress" explanations while giving greater control to the user with selections and mutations.

The representations used for the interactive explanations were primarily charts and texts. As shown in Figure 4.15 (Middle), tables were useful to support *Filter/focus* and *Compare* interactions. Textual explanations often came with *Clarify* interactions. Rules, although not appearing frequently in the corpus (5 times), were used to support *Clarify* and *Answer* interactions. Indeed, rules are easy objects for users to modify, create or delete, as exemplified in [Guo et al., 2022, Hepenstal et al., 2021, Ming et al., 2019].

Figure 4.15: Left: Frequency of the interaction categories used in the corpus and frequency of their combinations ; Middle: Percentage of studies using an explanation representation per interaction category; Right: Percentage of studies focusing on a type of user question per interaction category /

4.4.3 *Evaluating interactive explanations*

To address our RQ₄, we report below how XAI researchers in our corpus have been measuring explanations and explainable AI systems based on human-grounded evaluations [Doshi-Velez and Kim, 2017]. Below we provide brief descriptions of the measures and highlight trends and challenges in evaluating interactive explanations.

Few controlled experiments. Few empirical studies supported a cross-sectional analysis of results on interactive XAI by using a static explanation as a baseline. Most papers (20/48) did not use any control condition (*cf.* Figure 4.17). Even if the measures in these articles are sometimes quantitative as in [Hernandez-Bocanegra and Ziegler, 2021] where the authors measured different constructs (system efficiency, transparency...) on Likert scales from 1 to 5 points, these results are hard to interpret in comparison with the rest of the XAI literature.

Nine of the 48 articles in our corpus compared interactive and static explanations through between-subject experiments. These comparisons were very informative for analyzing the added value of interactivity in XAI. We provide in Section 4.4 a qualitative analysis of the added value of interactive explanations based on this work. To a lesser extent, comparisons between interactive explanations and no explanation (13/48 items) are also useful for understanding the benefit of interactive explanations. We also leveraged this body of work in Section 4.4. Other context-specific comparisons were made between an interactive explanation and other explanation types [Schaffer et al., 2015, Guo et al., 2022, Suresh et al., 2022, Wilkinson et al., 2021], other interactive systems [Polley et al., 2020, Yan et al., 2020], other AI models [Ross et al., 2021], other interactivity types [Ghai et al., 2021] or random baselines [Jia et al., 2022], among others. Some of these user-based evaluations were within-subject experiments [Dodge et al., 2022, Springer and Whittaker, 2019, Feng and Boyd-Graber, 2019].

Much of the work that did not use a baseline provided valuable qualitative assessments instead. This research often employs usage scenario (or "use cases") to study users' reactions to XAI systems in realistic settings [Kwon et al., 2019, Ming et al., 2019, Jia et al., 2022, Cheng et al., 2022]. These qualitative insights often focused on capturing the user's perceived ease of use and/or usefulness of the XAI system (16/20 papers).

A wide toolbox. We identified 19 different metrics to evaluate XAI systems with users from our corpus. Fourteen of them were used twice or more: perceived usability, perceived usefulness, understanding, perceived explanation length/quantity, time, trust, cognitive load, performance at task, learning, predicted accuracy, perceived control, perceived fairness, perceived transparency and reliance (*cf.* Figures 4.16 and 4.17). Other measures were used such as perceived feedback quality and difficulty [Guo et al., 2022], explanation persuasiveness and sufficiency [Hernandez-Bocanegra and Ziegler, 2021], number of interactions (clicks, etc.) with the explanations [Naiseh et al., 2021a] and naturalness and humanness

of the explanations [Rebanal et al., 2021]. Table 4.4.3 provides the definitions used for each of these metrics.

We recognized four of the five user-based measures for evaluating XAI systems outlined in [Hoffman et al., 2019]: user satisfaction, understanding, trust (and reliance) and human-XAI performance. Indeed, none of the papers in our corpus measured participants' curiosity, highlighting a gap in the literature for making XAI systems more engaging through users' feedback. However, we actually found more than five types of human-based metrics. Measures of the propensity of XAI systems to enhance learning, perceived transparency and fairness, humanness and naturalness of explanations, or cognitive workload, provide additional nuances to the XAI researchers' toolbox.

The many shades of user satisfaction. User satisfaction was the most frequently used measure in the corpus. However, we found many nuances of this concept. Some assessed whether users liked the systems [Kim, Chris et al., 2021, Jia et al., 2022], and/or found them useful [Jin et al., 2020, Khurana et al., 2021, Sovrano and Vitali, 2021], helpful [Yan et al., 2020, Jacobs et al., 2021], effective [Hernandez-Bocanegra and Ziegler, 2021] and/or easy to use [Szymanski et al., 2021, Kwon et al., 2019], or preferred the explanation or explanation system over another. In order to capture some of these nuances while keeping the papers coding manageable, we divided user satisfaction into two main clusters: ease of use (*i.e.*, perceived usability) and perceived usefulness of the XAI system.

Some articles already made distinctions between these two constructs [Jin et al., 2020, Szymanski et al., 2021], but others did not, especially when using questionnaires such as the Explanation Satisfaction Scale [Hoffman et al., 2019], which incorporates both usability and usefulness concepts [Bove et al., 2022, Guo et al., 2022]. When this was the case, we reported the measure under both "usability" and "usefulness".

Definition

Perceived usability. *A user-reported measure of how easy and likeable something is to use.*

Under the "perceived usability" construct, we included measures of usability, ease of use, likeability, *i.e.* whether users expressed that they liked the interactive explanation (or the XAI system)—typically through a one-item questionnaire [Guo et al., 2022] or through a qualitative think-aloud study [Jin et al., 2020],—and user preference, *i.e.* whether users preferred the system to a given baseline. Questionnaires such as the Post-Scenario Questionnaire [Lewis, 1991] or the User Engagement Scale [O'Brien et al., 2018] were often used to measure usability.

Definition

Perceived usefulness. *A user-reported measure of how useful something is to achieve the users' goals.*

In the concept of usefulness, we reported the accounts of "usefulness" and "perceived effectiveness", the latter being assessed through Tintarev's

questionnaire [Tsai et al., 2021, Hernandez-Bocanegra and Ziegler, 2021, Tintarev, 2007].

Joint use of subjective and objective measures. Many self-reported measures have an objective equivalent, and the papers in our corpus have taken advantage of this. This was the case for understanding, trust and cognitive load.

Understanding was most often measured subjectively by asking participants if they understood the system [Bove et al., 2022, Chromik et al., 2021]. However, some also assessed understanding objectively by asking carefully designed, often context-specific questions [Bove et al., 2022, Cheng et al., 2019, Ming et al., 2019, Rebanal et al., 2021]. Predicted accuracy, referring to the ability of users to predict what the system will output given certain entries, has been measured in [Nourani et al., 2021, Chromik et al., 2021, Springer and Whittaker, 2019] and could be considered, as some argue [Chromik et al., 2021], as an objective understanding of the system.

Participants' trust in the system or explanations was mostly assessed subjectively, by asking people to report their confidence in the XAI tool. McKnight's framework was used in three studies [Ghai et al., 2021, Wilkinson et al., 2021, Hernandez-Bocanegra and Ziegler, 2021]. Other papers referred to Tintarev's [Tintarev, 2007] measures of trust [Hernandez-Bocanegra and Ziegler, 2021, Wilkinson et al., 2021, Tsai et al., 2021]. [Hernandez-Bocanegra and Ziegler, 2021] also used items from Kouki et al. [2019] to measure trust related to explanations rather than to the system. However, trust was also measured objectively, by observing users' ability to reject an incorrect AI suggestion [Ribeiro et al., 2016, Liu et al., 2021, Buçinca et al., 2021, Kim, Chris et al., 2021]. We referred to this measure as "reliance", but [Kim, Chris et al., 2021] framed it more positively as "user skepticism", while others have called it "human-AI agreement" [Liu et al., 2021].

Users' cognitive workload when interacting with XAI systems was reported in five studies. It was measured by the NASA-TLX workload index, or a subset of its items. Closely related to cognitive load are estimates of the time spent on the XAI system or explanation, and the perceived length and/or complexity of the explanation. The former is an objective, quantitative estimate, while the latter is a self-reported measure [Kouvela et al., 2020, Buçinca et al., 2021, Szymanski et al., 2021].

The quality of self-reported measures can sometimes fall short of researchers' expectations, as some [Dodge et al., 2022, Naiseh et al., 2021a, Wang et al., 2019a] argue. Objective measures of understanding, trust and cognitive load may offer more reliable observations, even though at present, their measures are less standardized and more context-specific, making results more difficult to compare across different studies. Dodge et al. [2022] notably proposed "the ranking task" as an alternative to self-reported measures.

Task performance as the new benchmark. Some work [Buçinca et al., 2020, Bansal et al., 2021] advance that subjective measures could be misleading to properly assess the added value of explanations. Buçinca

et al. [2020] found that an increase in user satisfaction did not necessarily lead to improved performance, if not the opposite. Instead, Buçinca argues, measuring task performance should be the standard benchmark as it comes down to directly evaluating XAI systems against what they were designed for: increasing humans' autonomy and complementarity with AI. While XAI may serve other purposes, such as increasing user confidence and understanding, measuring task performance has the advantage of being a metric that is both objective and easily quantifiable. In fact, many empirical studies in the corpus have adopted it (21/48). Some articles also measured other constructs related to the task at hand, such as task complexity or time spent performing the task [Ross et al., 2021].

Less frequent goal-specific metrics. Evaluation measures are chosen in relation to the purpose that explanation serve. For example, Lee et al. [2019] and Anik and Bunt [2021] aimed at increasing public transparency and perceived fairness of an AI system. Therefore, Anik et al. used the questionnaire from [Binns et al., 2018] to assess users' perception of the fairness of the system and Lee et al. relied on their own quantitative metrics by asking participants to indicate on a Likert scale their agreement with the sentences "My assignment is fair", "This participant's assignment is fair", or "The overall group outcome was fair". Similarly, learning was a few times measured as a separate concept from the understanding of the AI model. Measures of "learning" focused on how well XAI explanations and systems helped users learn about a topic such as gender bias ([Melsión et al., 2021]) or self-care awareness ([Tsai et al., 2021]). In conversational interfaces, explanations were evaluated according to their humanness and engagingness [Hepenstal et al., 2021, See et al., 2019], to their persuasiveness [Hernandez-Bocanegra and Ziegler, 2021], or their naturalness [Rebanal et al., 2021].

4.4.4 *Interactive explanations increase trust, but not necessarily overtrust*

In Chapter 3 we found some evidence that explanations tend to increase trust, even when it is unwarranted. However, it is still uncertain whether interactivity in explainability can mitigate or resolve these problems by better matching human cognitive processes. While theoretical work in education and psychology outline the benefits of interaction for explanation and learning [Roussou, 2004, Miller, 2019], empirical results do not always align with these statements. In [Liu et al., 2021] for example, they find that interactivity could increase human biases and overreliance on AI. This subsection summarises the effects of interactive explanations on trust and reliance using the controlled and qualitative evaluations in our corpus. We base our qualitative findings on the summary presented in Figure 4.19, and on the qualitative analyses of the effects of interactivity provided in the corpus.

No clear indication of an interactivity effect on overtrust, overreliance or cognitive load. Some concern has been expressed that interactivity could increase users' cognitive load and their overreliance on AI [Liu

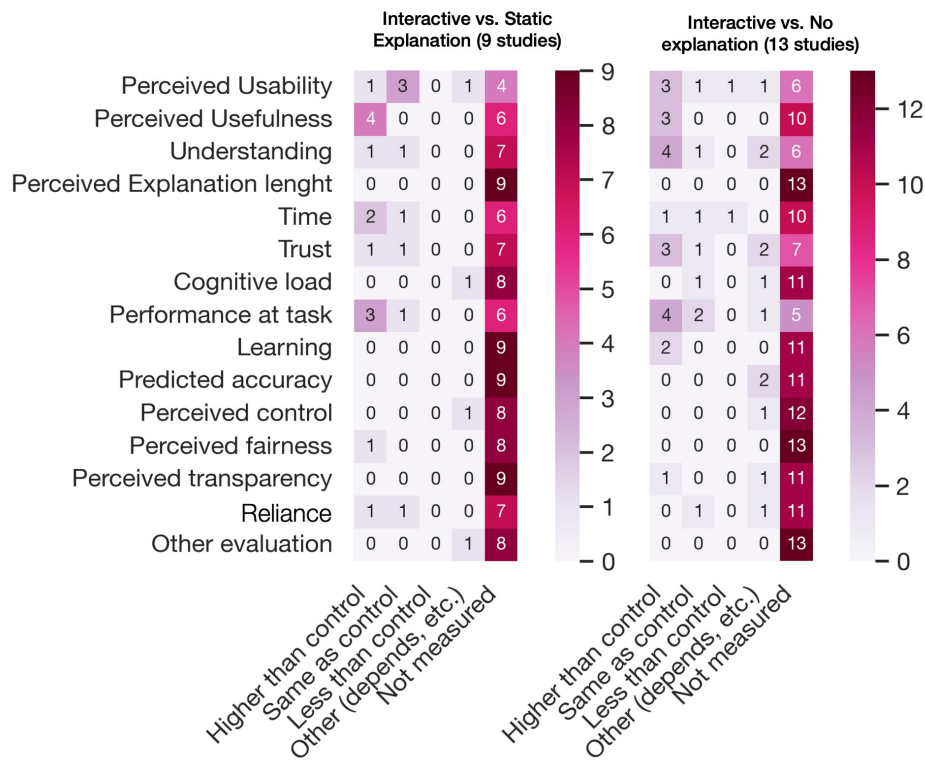


Figure 4.19: Left: Count of the positive, negative and neutral quantitative evaluations of interactive explanations compared to static ones, against various user-based metrics, based on 9 different studies. Right: Count of the different evaluation outcomes in the empirical studies comparing interactive explanations with no explanation as a baseline, extracted from 13 different papers in the corpus.

et al., 2021]. We did not find many results to either confirm or refute this. The results for user cognitive load were generally not directly related to explanations alone, but to other external factors, either with the static or no-explanation baseline. Buçinca et al. [2021] and Ghai et al. [2021] highlighted the importance of the user's individual need for cognition, knowledge of the task to perform, or of the model used [Ross et al., 2021]. Qualitative analyses suggest, however, that *Simulate* interactivity techniques can increase users' perceived difficulty of interacting with the system as we detail in the paragraph 4.4.5.

Compared to no explanation, interactive explanations did not lead users to over rely more on the AI. However, results were mixed for the comparison of interactive explanations to static ones. On the one hand, using *Simulate* interaction techniques, Liu et al. [2021] found that interactive explanations could increase users' tendency to blindly trust the AI. On the other hand, Buçinca et al. [2021] found that their on demand interactive features in the *Clarify* style could significantly decrease overreliance. The interactivity type therefore seems to be instrumental in the development of overreliance.

Higher perceived control leads to greater perceived fairness, perceived transparency, and (less clearly) trust. A participant in [Yan et al., 2020] said "I want to know why it is biased, not have the machine tell me why". This highlights the power of user controls and interactivity to drive trust and support users' autonomous exploration of the AI model. Lee et al. [2019] confirmed this with quantitative evidence, finding that *Reconfigure* interactions significantly improved perceived fairness. The

authors mentioned that the *Answer* interaction—here participants could correct the algorithmic allocation—caused users to perceive the model as fairer.

We did not find a substantial trend in the effect of interactivity on trust in the quantitative studies in the corpus. As indicated by the right side of Figure 4.19, the results in [Khurana et al., 2021] and [Cheng et al., 2019] do not converge. Some studies described the link between trust and external factors such as users' prior experience with AI [Ghai et al., 2021] or on users' individual propensity to trust [Kim, Chris et al., 2021].

4.4.5 *Interactive explanations are useful, but not easy to use*

To take stock on the benefits of interactivity in explainability, we present below a summary of empirical evaluations of interactive XAI on several user based metrics other than trust.

Interactive explanations improve perceived usefulness but not usability. Overall, there appears to be repeated evidence that interactivity does not significantly improve perceived usability [Guo et al., 2022, Sovrano and Vitali, 2021, Lee et al., 2019] compared to static explanations, but it does improve perceived usefulness [Bove et al., 2022, Ghai et al., 2021, Buçinca et al., 2021]. However, when compared to a baseline of no explanation, interactive explanations lead to an increase in perceived ease of use [Hepenstal et al., 2021, Tsai et al., 2021, Kim, Chris et al., 2021]. This reinforces the hypothesis that interactivity is not responsible for the improvement in perceived usability, but the presence of explanations is. It is possible that interactivity increases the complexity of the system, but at the same time supports users in their task and exploration of the models. The authors of the Gamut interface [Hohman et al., 2019] state that "interactivity was so fundamental for our participants' understanding of the models, that when we prompted them to comment on interactivity, people could not conceive non-interactive means to answer both their hypotheses and prepared questions". This study illustrates the potential of interactivity in terms of usefulness and as a factor in enabling users to achieve their goals.

Interactive explanations improve performances of the (human+AI) team, sometimes increasing time spent on explanations. Human+AI team performance was found to be improved in [Ghazimatin et al., 2021, Buçinca et al., 2021, Lee et al., 2019] with interactive versus static explanations. However, in two other studies [Cheng et al., 2019, Buçinca et al., 2021], the time spent to interact with the explanation system was higher for interactive explanations compared to static ones. The presence of interactive explanations compared to a "no explanation" baseline also improved task performance. These results seem logical, as greater interactivity can help users dive deeper into exploring a model and augment their cognitive engagement in the process. However, increasing the number of interactions with the system, as well as deeper analytical thinking, would understandably take more time. For example, interactivity can be designed to elicit user cognitive engagement such as in [Buçinca et al.,

2021], which in turn can enhance task performance. Further, Buçinca et al. [2021] showed that on demand explanations—from the *Clarify* interaction category—could significantly increase the performance of the human+AI team compared to static explanations.

However, Naiseh et al. [2021a] demonstrated that an interactive friction-based feature—falling in the *Answer* category—could lead participants to interact significantly more with the system, while having no impact on the time spent using the system.

Unclear role of interactivity on understanding and learning. From Figure 4.19, it appears clearly that the presence of (interactive) explanations compared to no explanation enhances user understanding of a model. Similarly, learning seems to be persistently enhanced by the presence of interactive explanations [Tsai et al., 2021, Melsión et al., 2021]. At the same time, user understanding of a model was dependent on other factors, including the order in which users saw weaknesses in the system [Nourani et al., 2021], or the stage of interaction with the system [Chromik et al., 2021], or the type of model that was explored [Ross et al., 2021]. In addition, Cheng et al. [2019] found that interactive explanations led to higher objective and subjective understanding of the model compared to a static baseline, but Bove et al. [2022] could not find any statistically significant improvement of interactive over static explanations for both objective and subjective understanding. More work is therefore needed to clarify the added value of interactive explanations over static explanations for understanding and learning.

Qualitative evidence of the added-value of a few interaction techniques. Despite the unclear quantitative evidence, the qualitative analysis of the corpus suggests that understanding is facilitated by interactivity. For example, one participant reported that receiving feedback and interacting with the model helped him "learn from my mistakes and expose my misconceptions" [Dodge et al., 2022]. Sevastjanova et al. [2021] showed that participants appreciated the on demand display of explanations as well as the ability to edit them. Morrison et al. [2018] emphasized the usability of *Compare* interactive features to support human cognitive processes, finding that "comparison is much easier than classification for a person". Schaffer et al. [2015] demonstrated qualitatively that linear interactivity was perceived as useful. Furthermore, Springer and Whittaker [2019] highlight the need for progressive disclosure of model information in order to prevent users from seeing their expectations violated and distrusting the system when it is correct.

"Simulate" interactions can strain users' memory and time. While interactive explanations of the type *Simulate* have been evaluated positively on many fronts, notably usability, usefulness and understanding, they also seem to take up more time as qualitative analyses in [Bove et al., 2022, Ghai et al., 2021] show. Additionally, after using a simulation-based interaction feature, a participant in [Jia et al., 2022] indicated that: "At the end of the design process, I think my brain is stuck. I do not know what I have specified before. When I want to add a new attribute, I need to

go back to check if I have specified it already". This calls for a careful consideration of the natural tendency of people to lose track of previous simulations in the design of *Simulate* interactions. Consistent with this observation, Ross et al. [2021] found that user performance in recreating an outcome through perturbations of concept-features degraded as the dimensionality of the concept-features increased. Future research should therefore design simulation explanations taking into account the limitations of people's memory.

Current dialogic explanations lack humanness. In [Rebanal et al., 2021], participants rated the naturalness of conversational explanations more harshly than the other measured aspects of the explanations.

Also, in [Tsai et al., 2021], participants reported a similar lack of naturalness for the questions that were asked by the system to the user. The authors describe: "our participants felt confused about the questions asked by the [conversational agent] in terms of the sequence, quantity, and relevance." However, in [Hepenstal et al., 2021] participants indicated they preferred to be able to "recognize when they were talking to a human or to a machine", actually preferring that humanness levels of explanations remain low. This questions the validity of aiming for more "dialogic" explanations that replicate a human-like explanation process. We provide more thoughts on this issue in the following section.

4.5 Discussion

We discuss below two open issues in interactive XAI. First, interactivity itself needs to be explained to users, adding another layer of complexity to XAI systems. Second, it is unclear whether dialogic/human-like explanations should be considered the ideal form of explanation communication by XAI researchers.

4.5.1 Interactivity calls for meta explanations

Interactivity itself requires some learning by the user [Roussou, 2004]. In addition to learning about the model, users must learn how to use the controls of the interface.

Hepenstal et al. [2021] observed that participants had many questions about how to use the interface and control it—"Can I click on that?". With *Answer* interactions, Tsai et al. [2021] also found that some participants felt confused by the questions asked by the system. They suggest that it would be helpful to provide additional explanations answering questions like "why does the system ask these questions?", or "how many questions would be asked or needed?" [Tsai et al., 2021]. These observations align with Sun et al. [2022]'s categorisation of user questions. One of them is called "Control", and is defined as "Questions about options for customizing or specifying preferences for how the model should work". Therefore, interactivity adds a layer of explanation in addition to model explanations.

We can make a parallel with the concept of meta-explanation introduced in [Dazeley et al., 2021]. Dazeley et al. [2021] point to a major issue in XAI research, which is the user's need to know where explanations come from in order to be able to trust the model and its explanations. As the authors put it: "if we cannot trust the agent's original decision, how can we trust the agent's explanation of that decision?". They call "meta-explanations" the explanations about the explanations themselves. Meta-explanations introduce a paradox whereby more explanations calls for more explanations, leading to unsustainable complexity. Similarly, explanations on the control of the interface could lead to cognitive overload and effects such as users ignoring explanations and AI predictions, as described in [Tsai et al., 2021].

Our corpus highlighted diverging results on whether interactivity has an effect on cognitive load. Our analysis highlighted, however, the role of individual factors to drive cognitive workload. There is therefore a need for future research to investigate how to tackle the meta explanation paradox in the context of interactive XAI, and how to find the right level of explanation for each user [Dazeley et al., 2021, Bućinca et al., 2021].

4.5.2 *Are dialogic explanations really the grail?*

According to Miller [2019] and Graaf and Malle [2017], people expect explanations to follow the conceptual framework of a social interaction. One reason for this is that people attribute human traits to XAI agents and therefore expect them to follow social conventions [Graaf and Malle, 2017]. Therefore, a good explanation would be provided through a social conversation. In fact, at least two studies from our corpus provided quantitative evidence that explanations communicated through *Ask* interactions improved perceived usability and understanding.

However, the participants in [Hepenstal et al., 2021] were bothered by the humanness of the XAI agent and preferred to have it made clear that they were not talking to a real person. Instead, they preferred robot-like explanations with "logical and clear responses". Indeed, while explainability should bring trust, anthropomorphism through human-like conversations can diminish trust by giving people the feeling of being manipulated. Hepenstal et al. [2021] suggest that different evaluation metrics could be applied to assess conversational XAI, such as understanding and bias mitigation, which are more representative of explainability's purpose.

If we take Miller [2019]'s depicted ideal of an AI agent's explanation⁴, perhaps a more important criteria than the social structure of the explanation would be the range of questions the explaining agent is able to answer. Overall, further theoretical work may be needed to clarify what "social interaction" means, whether it refers to its dialogue structure or to the social rules it abides by, such as Grice [1975]'s maxims. Future work could also examine the extent to which a "social" interaction with an AI agent can resemble human conversations, or even if this comparison makes sense.

⁴Miller presents it as a conversation, not necessarily in natural language, where the user asks a first request and follow-up questions

4.6 *Limitations*

One of the main limitations of scoping reviews is that they do not formally appraise the quality of the included studies [Arksey and O'Malley, 2005] through the means of, for example, the Cochrane Risk of Bias or other quality assessment tools. While this is compatible with the objectives of this survey—to identify, map and discuss evidence on empirical results in interactive XAI—we remind the reader again of this limitation.

Furthermore, although we applied a standardized methodology to identify articles, it is possible that relevant papers were missed because they were not published in peer-reviewed conferences or journals, because they were not present in the databases we surveyed or because they did not match our keyword search. This was the case for [Slack et al., 2022], which was published in a workshop and was therefore excluded during the eligibility phase, or for [Wu et al., 2021] which did not appear in the databases we searched. Indeed, as mentioned earlier, we chose to focus on HCI-oriented databases (ACM DL and IEEE Explore) rather purely AI ones, which may have led us to leave out relevant work in CS-focused venues. Since our interest is in interactivity and user studies, it seemed reasonable to limit ourselves to academic venues in HCI. Other work like [Krause et al., 2016] and [Kulesza et al., 2015] were not included in our study because the authors use the terms "interpreting" or "explanatory" in their title/abstract as references to the "explainability" notion. However, we believe that it would have been difficult to define the verbs interpret or explain and their conjugations as keywords because of their ubiquity. To remedy the limitation of a keyword search for the interactivity dimension, we searched for papers presenting an interactive XAI system in the eligibility phase instead of the identification phase [Moher et al., 2009]. This enabled us to include papers presenting interactive XAI solutions even though they did not express or emphasize in the abstract their contributions to the interactive explainability field.

In addition, we acknowledge that there may be a positive outcome bias [Callaham et al., 1998] in the results on interactivity because we searched published articles. We hope that by highlighting areas of uncertainty where it is unclear whether interactivity has positive or negative effects, this work will encourage others, including publishers, to consider all types of outcomes, including neutral or negative.

Then, although steps were taken to ensure consistency in our coding—including a final review of all the codings by one researcher—the final matrix may reflect each reviewer's own way of thinking.

Finally, it is possible that the summary of the papers' findings in Sections 4.4.5 and 4.4.4 may not capture the nuance of each context in which the results were found. However, it does provide a high-level, qualitative view of the results of empirical studies, and that was our goal.

4.7 Conclusion

This chapter presented a review of the literature on interactive explanations evaluated with human users. We provided a qualitative analysis of 48 papers shedding light on (1) the types of interactivity techniques that have been used so far in XAI, (2) the context in which interactive explanations were implemented, (3) the metrics used to evaluate interactive explanations with human users, and (4) the effects of interactivity on user satisfaction, understanding, trust, performance at task and other user-based metrics.

We provided a classification of XAI-specific interactivity techniques which can serve as a basis for explainability system designers to navigate the interactivity spectrum in XAI.

Our analysis showed that attention has been focused on interactivity that allows for input modification, but less attention has been paid to perturbing outcomes of AI systems, and to dialogic interactions. Combinations of dialogic interactions with interactions that allow mutation or selection is an under-explored area. The evaluation metrics we observed provide a wide range of ideas for XAI researchers to evaluate their systems against what they were designed for. Finally, we found converging results regarding the effect of interactive explanations on users. We identified that interactivity increases perceived usefulness and the performance of the human+AI team compared to static explanations, but it does not improve usability. In addition, it increases time spent by users on XAI systems. The empirical studies gathered in our corpus also demonstrated conflicting results on the role that interactivity has on overreliance, cognitive load, learning and understanding. This highlights grey areas to be addressed in future empirical research.

We hope that this work will help future research to share a common vocabulary on interactive XAI. Also, we hope it will facilitate future systematic reviews to identify best practices in interactive XAI design, as more empirical research is conducted in this area.

In the next part, we contribute to the ongoing efforts in explainability to test explanations' needs and effects empirically. We study explanation needs in two applications of AI in the financial sector, taking a human-centric approach.

Evaluation concept	Definition	Main evaluation methods
Perceived usability	User's perception of how easy to use the explanation user interface is.	Adapted question items from Explanation Satisfaction Scale [Hoffman et al., 2019], Post-Scenario Questionnaire [Lewis, 1991] or the User Engagement Scale [O'Brien et al., 2018]; qualitative think-aloud study [Jin et al., 2020].
Perceived usefulness	User's perception of how useful, effective or helpful the XAI system is for achieving their goals.	Question items from Tintarev's questionnaire [Tintarev, 2007], Explanation Satisfaction Scale [Hoffman et al., 2019] or [Vandenbosch and Ginzberg, 1996]; qualitative think-aloud study [Yan et al., 2020].
Understanding	The extent to which the user understands a model or its explanations.	"Objective understanding": Likert-type, context-specific questionnaires [Bove et al., 2022, Cheng et al., 2019, Ming et al., 2019, Rebanal et al., 2021], "Subjective understanding": qualitative think aloud or free-text analyses, e.g. [Bove et al., 2022, Chromik et al., 2021].
Perceived explanation length/quantity	User's perception of the length or quantity of the explanation, often used as proxies for the complexity of the explanation.	Direct questions about the quantity, length, or complexity of the explanation, e.g. [Kouvela et al., 2020, Buçinca et al., 2021, Szymanski et al., 2021].
Time	The time spent by the user interacting with the XAI system to perform a task.	Direct measure of the interaction time, e.g. [Ross et al., 2021].
Trust	User's willingness to depend on an XAI system because of the characteristics of the system [Mcknight et al., 2011, Rousseau et al., 1998].	Question items from McKnight's framework [Mcknight et al., 2011], Tintarev's questionnaire [Tintarev, 2007] or Kouki et al. [2019]'s measure of trust towards explanations.
Cognitive load	The amount of working memory resources used by the user while interacting with the XAI system [Miyake and Shah, 1999].	NASA-TLX workload index.
Performance at task	The performance of the human+XAI team in performing a specific task.	Measured through case-by-case metrics adapted to a context-specific task, e.g. [Dodge et al., 2022, Buçinca et al., 2021, Feng and Boyd-Graber, 2019].
Learning	How well explanations and/or XAI systems help users learn about a specific topic.	Context-specific questions usually defined by the authors themselves about a topic. See examples for learning about gender bias ([Melsión et al., 2021]) or self-care awareness ([Tsai et al., 2021]).
Predicted accuracy	User's ability to correctly anticipate the AI's behavior.	Number of correct guesses of the AI's prediction by the user [Nourani et al., 2021, Chromik et al., 2021, Springer and Whittaker, 2019].
Perceived control	User's perception of their control over the XAI system.	Adapted question items from the Knijnenburg et al. [2012] framework.
Perceived fairness	The extent to which users perceive the XAI system to be fair and transparent.	Fairness questionnaires from Binns et al. [2018] or Lee et al. [2019].
Perceived transparency	User's perceived understanding of the recommendation rationale	Adapted question items from Millecamp et al. [2019] or Tintarev [2007] frameworks.
Reliance	User's ability to reject an incorrect AI suggestion.	Precision and/or recall in correct rejections or acceptances of a prediction, e.g. [Ribeiro et al., 2016, Liu et al., 2021, Buçinca et al., 2021, Kim, Chris et al., 2021].

Figure 4.18: Evaluation concepts used twice or more in the corpus with corresponding definitions and evaluation methods.

PART II

*Complying with regulation
using human-centric explainable
AI: two case studies in finance*

Chapter 5: Empowering customers of robo-advisors with explainability presents a mixed-methods experiment (qualitative and quantitative) on the impact of different formats of explanations on customers' trust and empowerment in life-insurance underwriting. This chapter builds on the reflections presented in a 2022 workshop paper and on subsequent studies that were published as a conference paper in 2023:

"Towards Informed Decision-making: Triggering Curiosity in Explanations to Non-expert Users", Astrid Bertrand, 2022 Workshop on XAI and HCI, IHM Conference, Namur, Belgium, 2022 <https://hal.science/hal-03651368/document>.

"Questioning the ability of feature-based explanations to empower non-experts in robo-advised financial decision-making", Astrid Bertrand, James R. Eagan, Winston Maxwell, Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), Chicago, USA, 2023 <https://doi.org/10.1145/3593013.3594053>.

As the first author of these studies, I delineated the motivation and research questions with the guidance of my colleagues at the ACPR, notably Olivier Fliche and Christine Saidani, and both co-authors. I conducted interviews with supervisors and novice users, coded a fictitious robo-advisor using python and javascript, designed and coded explanation prototypes, conducted and analyzed the quantitative study, and wrote the paper. The methods, results, and text were discussed with all three co-authors.

Chapter 6: Understanding the supervisors' needs for explainable AI in financial crime detection presents a qualitative, mixed-methods analysis (leveraging HCI and legal approaches) of the perspective of regulatory supervisors on the role of explainability in the field of anti-money laundering. This chapter will soon be published as a conference paper:

"AI is Entering Regulated Territory: Understanding the Supervisors' Perspective on Model Justifiability in Financial Crime Detection", Astrid Bertrand, James R. Eagan, Winston Maxwell, Joshua Brand, was conditionally accepted for publication in the proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24), Honolulu, Hawaiï, USA, 2024.

As the first author, I delineated the motivation and research questions, designed and conducted all the workshops and interviews, and wrote the paper. The fourth co-author helped in the analysis of a few workshop transcripts. The methods, results, and text were discussed with all co-authors.

Chapter 5

Empowering customers of robo-advisors with explainability

THIS CHAPTER sheds light on the challenge of using algorithmic explanations for user empowerment and customer protection compliance. We examine in a real world scenario the "**explanation paradox**": on the one hand, explanations are necessary to inform users of critical information regarding the decisions made about them. On the other hand, Chapter 3 revealed that explanations tend to reinforce trust, even when it is unwarranted, making customers more vulnerable to inappropriate recommendations. In this chapter, we therefore explore the potential of human-centric explainable AI to address this challenge.

Specifically, we investigate whether legally required feature-based explanations for life-insurance robo-advisors¹ help clients make better financial decisions. We also consider the perspective of regulatory supervisors in customer protection in life insurance. We find that providing feature-based explanations does not improve appropriate reliance or understanding compared to not providing any explanation. In addition, dialogic explanations increase users' trust in the recommendations of the robo-advisor, sometimes to the users' detriment. This real-world scenario illustrates how XAI can address information asymmetry in complex areas such as finance. This case study was made possible by our collaboration with the ACPR², the regulatory authority for financial services in France.

We begin by presenting some background on the literature on XAI for non-expert users and on the context of life-insurance in Section 5.2. We then build Robex, an explainable robo-advisor, to enable our domain-driven, contextual enquiry, using market research. We design explanations of Robex using co-design with end-users and regulatory supervisors. We present the methodology for this co-design qualitative study in Section 5.3. We redesign our explainability prototype based the needs of non-expert clients and the requirements of regulatory supervisors, experts in customer protection, in Section 5.4. In a subsequent study, we use Robex to quantitatively compare the effectiveness of various explanation formats in helping users understand, and appropriately rely on recommendations. We test the capacity of explanations to meet the cus-

¹ Robo-advisors are online platforms that provide financial advice.

² In French "Autorité de Contrôle Prudenciel et de Résolution"

customer protection objectives pursued by financial regulation. We present the methodology used for this quantitative experiment in Section 5.5, and its results in Section 5.6. Section 5.7 discusses the implications of our findings on the role of explainability to inform customers in finance.

5.1 Motivation and research questions

With the rise of commercial recommender systems, online AI-based services are becoming increasingly common. As a result, internet users are frequently presented with opaque personalized suggestions. While explanations are often unnecessary or non-critical in many low-risk applications of AI, such as for movie or music suggestions, they can be required by law in some high-stakes industries, such as finance. This is the case for systems distributing life insurance proposals in France.

Robo-advisors are democratising access to investing by enabling full online distribution of life insurance contracts and other investment plans. After answering a few profiling questions, users receive a recommendation for a life insurance contract that matches their financial situation. In recent years, these recommender systems have started touting AI to make more targeted suggestions. In Europe, financial legislation requires that the reasons for recommending a life insurance plan be explained according to the characteristics of the client, in order to empower her in making a "fully informed decision". In this context, the financial regulation aims at protecting clients from recommendations misaligned with their objectives, risk appetite and other personal characteristics.

Additionally, the forthcoming AI Act classifies³ AI-based robo-advisors as "high-risk"⁴, subjecting them to a demanding certification process and high transparency requirements in the near future.

Moreover, the financial domain can feel overwhelming and complex to many people [Prawitz et al., 2006], which poses an additional challenge: explaining in simple terms not only the attributes of the system but also financial principles to novice users. Few studies [Bibal et al., 2021] have focused on how to design legally required explanations for lay users in complex, high-stakes scenarios. As seen in Chapter 2, , cross-disciplinary efforts in both law and HCI are rare, and the regulatory challenges associated with explainability have not been fully explored by HCI researchers.

Nevertheless, recent advances in the fast-growing field of explainability have brought a better understanding of how different representations and interactions of AI explanations impact non-expert⁵ users [Szymanski et al., 2021, Bove et al., 2022, Cheng et al., 2019, Rebanal et al., 2021, Mohseni et al., 2021b]. Szymanski et al. [2021] found that lay users preferred graphical explanations but could more easily misinterpret them compared to textual explanations, motivating the need for hybrid textual and visual explanations. However, little is known about where the cursor should be placed between textual and visual content.

We aim to address these gaps by leveraging the knowledge of customer protection specialists. We believe the insights from experts from the regulatory sphere present interesting yet so far unsolicited proxies for characterizing the users' needs. We address the question of enabling warranted customer trust in recommender systems [Buçinca et al., 2021], which ties in with the research in the previous chapters.

Our research questions are as follows:

³ as of December 2023, based on the European Commission's proposal and the Council and Parliament's adopted texts.

⁴ Text adopted by the Council in Nov. 2022, Annex III, point 5: "AI systems intended to be used for risk assessment and pricing in relation to natural persons in the case of life and health insurance with the exception of AI systems put into service by providers that are micro and small-sized enterprises."

⁵ Here, "non-expert" refers to users who are either inexperienced in the domain task or inexperienced in using AI systems.

RQ1: *What are the regulatory expectations for explanations in financial investment services to protect customers? How can current XAI methods meet them?*

RQ2: *How do regulatory supervisors on the one hand and end users on the other describe the need for explanations?*

RQ3: *How effective are different representations of hybrid textual and graphical explanations to protect non-expert users?*

Our case study in life-insurance has implications for other profiling AI systems that interact with customers and data-subjects. For example, for systems making automatic individual decisions based on profiling, the GDPR requires to provide explanations such as "meaningful information about the logic involved"⁶.

⁶ Article 15(1)(h) General Data Protection Regulation (GDPR).

5.2 Background

This study falls in the HCI line of research on understanding explainability needs [Sun et al., 2022, Liao et al., 2020, Lim and Dey, 2009], and on testing explanations' effects with real users ("application-grounded evaluations" [Doshi-Velez and Kim, 2017]). We describe those research trends in Section 2.4 of Chapter 2. Specifically, we build on explainability research focusing on non-expert users. We highlight relevant findings below.

5.2.1 Mitigating overreliance issues for non experts

As reviewed in Chapters 3 and 4, some user studies evaluated the ability of XAI methods to successfully convey accurate mental models of AI systems to users. This line of research sheds light on the limitations of some technical solutions for aiding user understanding, or worse, on their potential for deception [Kumar et al., 2020, Kim et al., 2016, Ribeiro et al., 2016]. In Chapter 3, we found that user expertise, knowledge and skills appeared to be an essential factor for appropriate trust calibration in explainable AI systems. Specifically, non-expert users were more likely to be convinced by the mere presence of an explanation [Eiband et al., 2021, Fürnkranz et al., 2020, Lai and Tan, 2019], or to fell into confirmation or completeness bias [Szymanski et al., 2021]. Further, Simkute et al. [2020] stressed the importance of differentiating the reasoning of experts from that of lay users and reflecting this difference in the design of explanations. Quite logically, experts are able to be more critical of the explanations, sometimes at the cost of not trusting them enough, whereas lay users are more subject to overreliance [Schaffer et al., 2019, Bayer et al., 2021]. Explanations must therefore support either trust building for experts, or critical thinking for lay users.

Another key difference is the level of motivation to use explanations, which can be much lower for non-expert users. This makes it particularly challenging to make explanations both simple and appealing to lay users, while encouraging cognitive engagement and skepticism [Bertrand et al.,

2022, Naiseh et al., 2021a]. It is still unclear if explanations for non-expert users can be designed to foster trust and understanding while encouraging users' critical thinking (*i.e.* ability to detect errors) on the other. This may be desirable in sensitive contexts where algorithmic predictions may have a strong impact on the user's quality of life.

5.2.2 *Designing visualisations of AI explanations for non-expert users*

Some work has focused on the implementation of explanations for non-expert users in specific contexts [Szymanski et al., 2021, Bove et al., 2022, Cheng et al., 2019].

Cheng et al. [2019] presented explanations of an algorithmic school admission decision process to users with no domain or technical expertise. They found that static and interactive explanations, where users could change the inputs to see the resulting outcome, improved users' understanding of the AI decisions. Bove et al. [2022], however, were unable to replicate these results in the context of explaining an algorithmic car insurance pricing decision. They did not find that explanations improved comprehension but they did improve user satisfaction. Szymanski et al. [2021] studied how different representations of explanations, either visual, textual or both, affect users' understanding of an AI system in an artificial task⁷. The paper shows that purely visual explanations⁸ can be subject to misinterpretation, while purely textual explanations are better understood but less satisfactory to users. A combination of the two representations could therefore provide the best of both worlds. However, there may be many different ways to design "hybrid" textual and visual explanations. Additionally, it is still unclear if textual explanations presented as conversations achieve better user preferences and improve task accuracy compared to graphical formats.

⁷In the experiment, participants were tasked with estimating the reading time of news articles.

⁸in this case, line graphs

Then, explanations' ability to engage users in a sensitive and complex topic such as financial investment has not yet been studied in the XAI literature where artificial contexts are often used as test benches [Bućinca et al., 2021, Dodge et al., 2022, Feng and Boyd-Graber, 2019].

5.2.3 *Context: life-insurance distribution with "robo-advisors"*

In this chapter, we focus on a real-case application of explainability: explanations of online recommendations for life insurance products. In Europe, explanations in this context are legally required by sector-specific regulations to ensure customer protection. We describe below the case study context and the related legal requirements for explanations.

Overview. As AI systems gain performance, their adoption expands to areas considered critical. In finance, increasingly sophisticated recommender systems known as "robo-advisors" are democratizing online distribution of life insurance. In France, where the study was conducted, life insurance is a savings vehicle used both to pass on money to a designated beneficiary upon the death of the subscriber of the contract, and to make a long-term financial investment in a tax-advantaged environ-

ment. In the rest of the paper, we will only address the latter, most common usage of life-insurance. Life insurance subscribers are presented with a financial recommendation with a specific level of risk (a higher level of risk means more chances to win big but also more chances to lose). Choosing a life insurance contract with an appropriate risk level—not too high for the client’s financial situation—is crucial to ensuring clients’ financial stability. However, many clients may not be financially literate. Therefore, French and European legislation⁹ require insurance providers to produce “*clear, precise and non-misleading*” explanations to guide potential customers towards an “informed” decision and address the asymmetry of information between client and advisor. Most existing online recommender systems currently fall short of this explanation requirement, according to our discussions with French supervisors in the life-insurance sector. Specifically, explanations of online recommender systems, *i.e.* robo-advisors, rarely focus on the reasons why a recommendation is adapted to the user’s need, which is the type of explanation we focus on in this paper.

A trend towards more digital, AI-powered robo-advisors. The automated advice provided by robo-advisors is seen as a more cost-effective way of delivering propositions to parts of the population that otherwise have no access to financial advice, as highlighted in an OECD report [Mamiko, 2020]. In addition, the COVID crisis has accelerated the interest in online systems by increasing the demand for online and real-time services [Balasubramanian et al., 2021]. In France, most current robo-advisors are rule-based, with varying degrees of complexity in the amount and nature of the rules¹⁰. Yet, many studies foresee an acceleration of AI-based solutions to distribute financial services and in life-insurance plans [Balasubramanian et al., 2021, Mamiko, 2020]. AI-powered systems offer faster and more personalized financial advice. For brokers, data-driven profiling helps identify risk in a more fine-grained manner [Balasubramanian et al., 2020]. The insurance market is also gaining interest in AI-powered robo-advisors with the successful examples of companies which used this technology to increase sales revenue significantly [Balasubramanian et al., 2020].

Regulatory requirements for feature-based explanations. In the life-insurance context, financial legislation regarding the insurance sector apply. The law on insurance distribution (Articles 20 and 30 of Directive (EU) 2016/97 of January 20, 2016), which aims to protect consumers against the sale of products unsuited to their needs, specifies: “*The distributor shall advise on a contract that is consistent with the requirements and needs of the prospective subscriber and shall specify the reasons motivating this advice.*”¹¹. The text also mentions that: “*the distributor specifies in writing [...] the client’s requirements and needs and provides objective information on the insurance product offered in a comprehensible, accurate and non-misleading form to enable the prospective subscriber to make a fully informed decision.*” Further, the duty of information and advice in life insurance (L.522-5 of the French Insurance Code) requires to “*formalize the reasons for the appropriateness of the proposed contract in relation to the requirements and needs*

⁹The European Parliament and the European Council. 2016. Directive (EU) 2016/97 on insurance distribution.

¹⁰This was pointed out by the participants in our study who are supervisors of the life insurance sector.

¹¹Article L. 521-4 of the French Insurance Code

expressed.", which implies a requirement for feature-based explanations.

This leads us to question more precisely the purpose of the explanation in light of the objectives of the law. What exactly is expected of the explanation so that it is effective with regard to the objectives of the Articles L. 521-4 and L. 522-5 of the French Insurance Code and EU Directive 2016/97? One of the objectives of the explanations is to enable future life-insurance subscribers to make a "fully informed" decision about the product being proposed. This objective is explicitly stated in the text of Article L. 521-4 of the French Insurance Code and Article 20 of EU Directive 2016/97. However, this objective is relatively imprecise and difficult to measure. To better assess whether an explanation allows for an "informed" decision, the goal should be broken down into subgoals that are easier to verify. We understand these subgoals to be 1) help users appropriately rely on a recommendation (and be able to detect a big mistake) 2) help users understand a recommendation and why it is appropriate for them 3) help users calibrate their trust in robo-advisors. This is what we measured in Study 2.

In addition to the goal of "fully informing" clients, the law aims at enhancing the accountability of intermediaries by imposing the obligation to set out in writing the client's needs as well as the reasons why the recommended product is in line with those needs. The formalization of these steps will reduce the risks of intermediaries letting conflicts of interest interfere with their duty to give objective investment advice to customers.

In other contexts, AI systems may also be affected by requirements for feature-based explanations. Consumer protection law has provisions regarding explanations of recommender systems in online marketplaces. It notably imposes to show *"the main parameters determining the ranking [...] of offers presented to the consumer as a result of the search query and the relative importance of those parameters as opposed to other parameters"*¹². The General Data Protection Regulation [European Parliament and Council, 2016] provisions also apply in the case of entirely automated individual decisions based on profiling. It requires that data controllers disclose *"meaningful information about the logic involved"* (articles 13-15). The GDPR provisions apply *"when the decisions (i) involve the processing of personal data, (ii) are based solely on an automated processing of data and (iii) produce legal or significant effects on the recipient of the decision"* [Bibal et al., 2021, European Parliament and Council, 2016].

¹² New art. 6(a) of Directive 2011/83 on Consumer Rights

5.3 *Study 1 Methodology: a market-driven co-design approach*

5.3.1 *System design: Robex, the robo-advisor*

Robex¹³ is a simplified and fictional life-insurance recommender system developed for the purpose of this study. The recommendation algorithm of Robex is not AI but a rule-based algorithm established with the help of 4 domain experts, more precisely supervisors of the life-insurance industry. Indeed, since our goal was to study explanation representations using existing agnostic explainability methods, we did not need to use a real AI algorithm for this study. Similarly, the design of Robex was not our focus. However, we wanted our fictional robo-advisor to replicate the type of interface that robo-advisor clients would face. Therefore, we conducted a market analysis of existing online robo-advisors in France. This led us to review the design of four major players in France: Yonomi, Nalo, Linxea and Wesave¹⁴. For each of the identified robo-advisors, we tested the user journey from the profiling questionnaire to the simulation of the robo-advisor's proposal. We took inspiration from their content and interface design. This also allowed us to identify the classical steps in a robo-advisor user journey.

¹³ Standing for EXplainable ROBo-advisor

¹⁴ <https://www.yonomi.fr/>, <https://www.nalo.fr/>, <https://www.linxea.com/>, <https://www.wesave.fr/>

The usual subscription process with robo-advisors is as follows. First, users go through a series of questions about their profile and financial objectives. Then, they can see the summary of their profile and the proposed recommendation, on the same page. Robex follows the same first stages. During the recommendation phase, Robex presents an additional section on why this product is recommended to you.

The following elements from existing robo-advisors on the market have inspired us to implement similar features in Robex:

- the questions used in the profiling questionnaire about the user's characteristics (risk appetite, financial knowledge) and project. The ones we used in Robex are presented in Table 5.1.
- the brief, textual explanations in the profiling questionnaire to give some context and to indicate the answer to a question testing financial knowledge, as shown in Figure 5.4.
- the vocabulary used, driven by domain specificity and also by an intention to be accessible to all.
- the seamless navigation between the different steps of the user journey, and the clear presentation of the different stages upfront thanks to a progress bar at the top of the page, usually including "project", "simulation", "subscription", "documents", "signature". As with real robo-advisors, Robex presents a user journey progress bar, but adapted to the journey of the participants in our experiments.
- the presentation of the allocation of assets into large themes "actions", "obligations"... or by geographical region. However, all financial support and allocations in Robex were fictional.

- matching the user to one of a number of proposals with different levels of risk. Most robo-advisors propose a range of seven to ten proposals (which they sometime present as "user profiles"). We limited the range of proposals to five to reduce the complexity of our study.

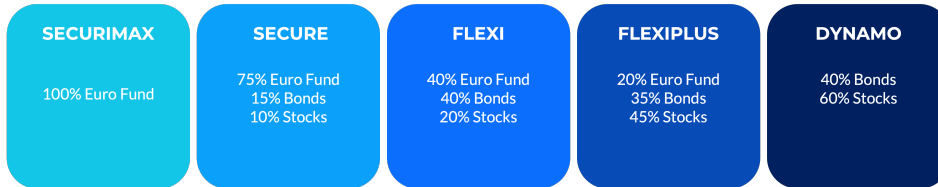


Figure 5.1: Fictional life-insurance plans proposed by Robex, the explainable robo-advisor developed for this study

In parallel, we conducted informal interviews with 4 supervisors with experience in the supervision of life-insurance distributors to better understand the domain. These discussions were instrumental in developing our own, simplified, profiling questionnaire to measure 5 user characteristics: the amount to be invested compared to the user's total financial wealth, her investment objective, her financial knowledge and experience, her risk appetite and the proportion of her financial assets already placed on financial markets. For each of the questions used to measure these characteristics (cf. Table 5.1), we associated coefficients so as to obtain a risk-score that denoted the amount of risk a user can take. We then sketched five fictional but realistic life-insurance plans that represent 5 levels of risk, as shown in Figure 5.1. Our score-based rules for insurance distribution then matched a profile to a plan. Robex is simplified because we have not taken into account the fees, investment horizons or performance of the funds in order to keep the complexity of the experiment manageable. The simplified Robex algorithm is presented in the Appendix B2.

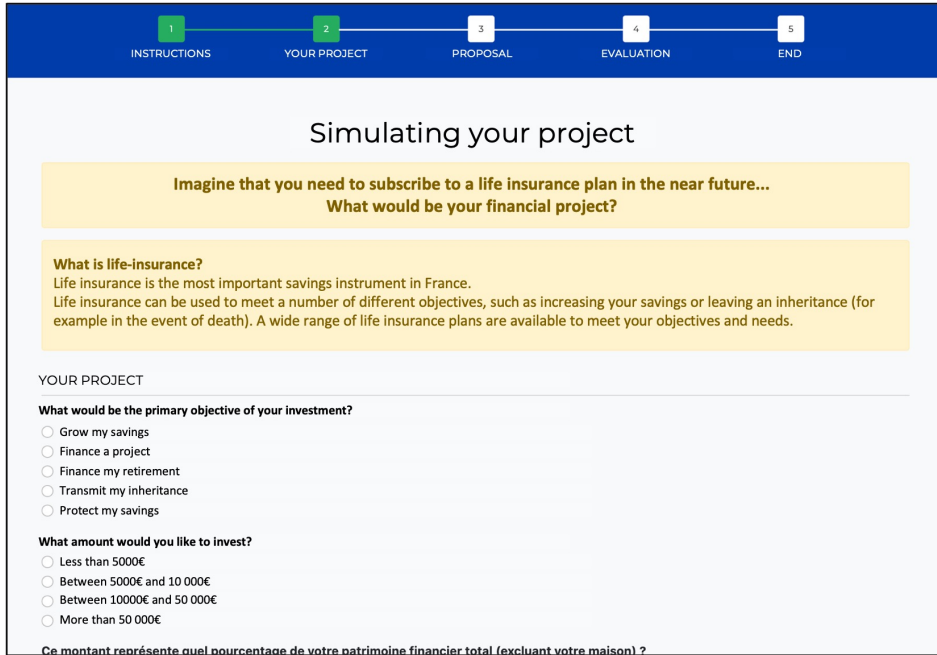


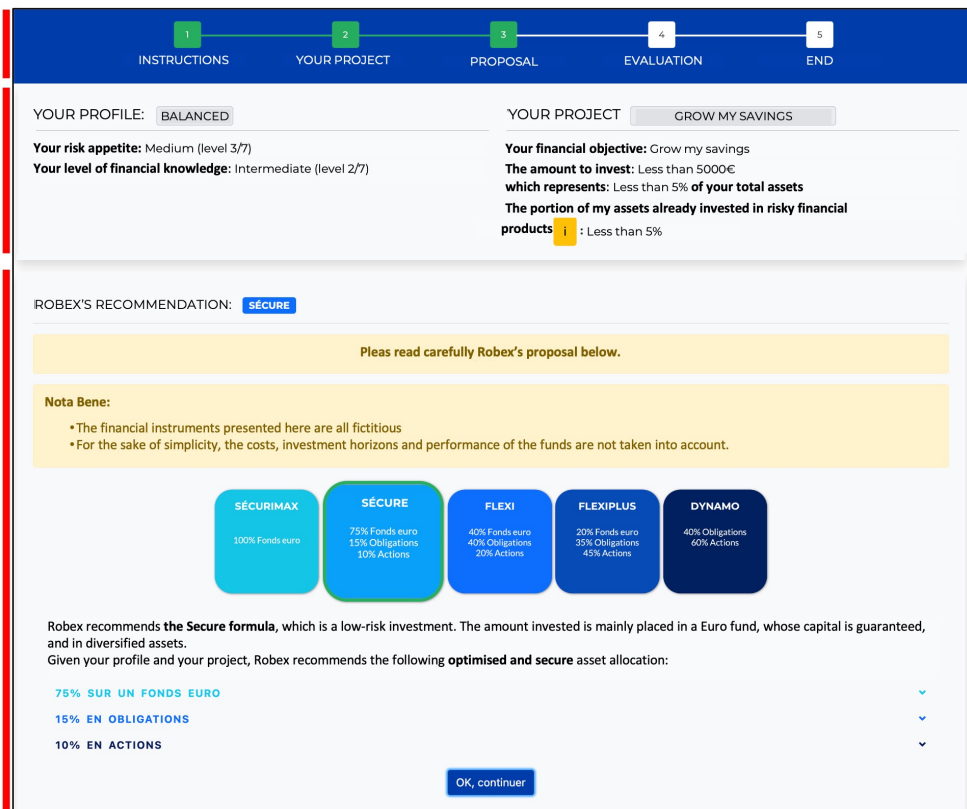
Figure 5.2: Screenshot of the Robex interface, showing the profiling questionnaire stage at the start of the user journey. Translated from French to English.

Figure 5.3: Screenshot of the Robex interface, showing the recommendation stage. As required by law, a summary of the user's profile is displayed first, followed by a life-insurance contract proposal with details. The explanation is presented on the same page, just after the proposal.

Study participant journey progression

Summary of customer profile and needs

Presentation of the robo-advisor's proposal



User charact.	Questions with [possible answers]
<i>Objective</i>	What would be the main objective of your investment? [Make my savings grow, Finance a project, Finance my retirement, Pass on my assets, Protect my savings]
<i>Amount to be invested</i>	How much would you like to invest? [Less than 5000€, Between 5000€ and 10 000€, Between 10000€ and 50000€, More than 50000€] This amount represents what percentage of your total financial assets (excluding your home)? [Less than 5%, Between 5% and 25%, Between 25% and 50%, Between 50% and 75%, More than 75%]
<i>Percentage of assets already invested</i>	Have you already invested in a financial product with a risk of capital loss? If so, how much of your total financial assets do these financial products represent? [Less than 5%, Between 5% and 25%, Between 25% and 50%, Between 50% and 75%, More than 75%]
<i>Risk appetite</i>	Which of the following statements is closest to the level of financial risk you are willing to take when saving or investing? [Take significant financial risk hoping for significant returns, Take above average financial risk hoping for above average returns, Take average financial risk hoping for average returns, I do not wish to take any financial risk] <i>For the next three sentences, please indicate the likelihood that you would engage in the specified behavior if you were in the situation described</i> "Investing 10% of your annual income in an investment consisting of securities issued by the European Union" [Very unlikely, Somewhat unlikely, Uncertain, Somewhat likely, Very likely] "Investing 5% of your annual income in highly speculative securities" [Very unlikely, Somewhat unlikely, Uncertain, Somewhat likely, Very likely] "Investing 10% of your annual income in a new business" [Very unlikely, Somewhat unlikely, Uncertain, Somewhat likely, Very likely]
<i>Financial</i>	Have you ever subscribed to a life insurance contract? [Yes, No]
<i>knowledge and experience</i>	Have you ever invested in a financial product with a risk of capital loss (e.g. PEA (Plan d'Épargne en Actions), multi-support life insurance contract, securities account, crypto assets, investment funds...)? [Yes, No] A high expectation of gains implies a high risk of capital loss. [True, False] A real estate fund (SCPI or OPCI) is a fund with guaranteed capital. [True, False] The capital invested in a life insurance plan is blocked for 8 years. [True, False] The capital invested in life insurance units of account is subject to a risk of capital loss. [True, False]

Table 5.1: Question used in the Robex's profiling questionnaire for measuring users' personal characteristics (translated from French to English).

Une espérance de gain élevée implique un risque de perte en capital fort.
 Vrai Faux Je ne sais pas

Vrai. Il n'existe pas de rendement élevé garanti. Autrement dit, une possibilité de rendement élevé s'accompagne toujours d'un risque élevé. (source : amf-france.org)

Un fonds immobilier (SCPI ou OPCI) est un fonds dont le capital est garanti.
 Vrai Faux Je ne sais pas

*Faux. Les fonds immobiliers, appelés SCPI (société civile de placement immobilier) ou OPCI (organisme de placement collectif en immobilier), sont des supports **présentant un risque de perte en capital**. Autrement dit, il n'existe pas de protection (ou garantie) du montant qui y est investi.*

Le capital placé en unités de compte d'une assurance-vie est soumis à un risque de perte en capital.
 Vrai Faux Je ne sais pas

Vrai. Dans un contrat d'assurance-vie dit "multi-support", il est possible d'investir une partie du montant sur des "unités de compte". Une unité de compte est un support d'investissement comme des actions, obligations, OPC (Sicav, FCP...), parts de SCPI, OPCI... Le capital placé sur des unités de compte n'est pas garanti.

Le capital investi sur un contrat d'assurance-vie est bloqué pendant 8 ans.
 Vrai Faux Je ne sais pas

Faux. Après un délai de 8 ans à compter de la date d'ouverture du contrat, l'assurance vie prévoit des conditions de retrait plus avantageuses. Cependant, l'assuré peut débloquer ses fonds quand il le souhaite, notamment avant 8 ans. Il s'expose juste à une fiscalité moins avantageuse. (source : amf-france.org)

Figure 5.4: Screenshot of the Robex interface, showing the answers it provided for test questions on participants' financial knowledge.

5.3.2 *Explanation prototype*

As seen in Section 5.2.3, the required explanations in life-insurance should link client's characteristics to the recommendation, which is what feature importance techniques do. To investigate the impact of feature importance explanations on users' trust and appropriate reliance on recommendations, we developed feature importance explanations in Robex.

We approached the explainability phase as if the rule-based recommender algorithm in Robex was a black-box. Our results can therefore be transposed to more opaque AI-powered robo-advisors. In each of the studies presented below, we used SHAP [Lundberg and Lee, 2017] a post-hoc, agnostic, and widespread interpretability method, to generate feature weights. We then use these weights as a basis for designing explanations that differ in representation format and interactivity.

One of our early prototypes is shown in Figure 5.5. We first designed the explanation interface taking inspiration from the graphical Shapley explanations presented in [Lundberg and Lee, 2017]. However, we tried to simplify the visual elements to make them readable by non-professional users. Specifically, we simplified the graph into a table, because some research on explainability showed that tables were the most interpretable representation medium for non-professional users [Huysmans et al., 2011]. The table sorts features per their influence on the risk of the prediction: features that decreased the risk of the proposal are shown in the left column and features that increased it on the right. We also applied a *card-based design* for the display of each feature-related explanation. This design enables to provide more context with each feature. Each card contains the name of the feature in boldface, the value of the feature for the user in grey, and its impact in natural language sentence [Bove et al., 2022].

We showed to participants in Study 1 a prototypical "graphical" summary of the importance of each variable on the risk of the proposal, as shown in Figure 5.5. We improved the explanation representation based on the feedback from expert and lay participants of the co-design experiment we present in the following section.

5.3.3 *Co-design sessions and analysis*

To answer our RQ1 and RQ2, we interviewed domain experts and lay users to better understand end users and supervisors needs and expectations, following a participatory design approach [Spinuzzi, 2005]

Procedure. Each participant took part in an individual session that lasted between 45 minutes and 1h30. The aim of the interviews was to collect users' feedback on our prototype, and work with users and domain experts to create explanations that meet their needs and requirements. This participatory design approach has already been endorsed in the field of explainability, for example in [Panigutti et al., 2023a, Cheng et al., 2022, Wang et al., 2019a]. Each co-design session was divided into

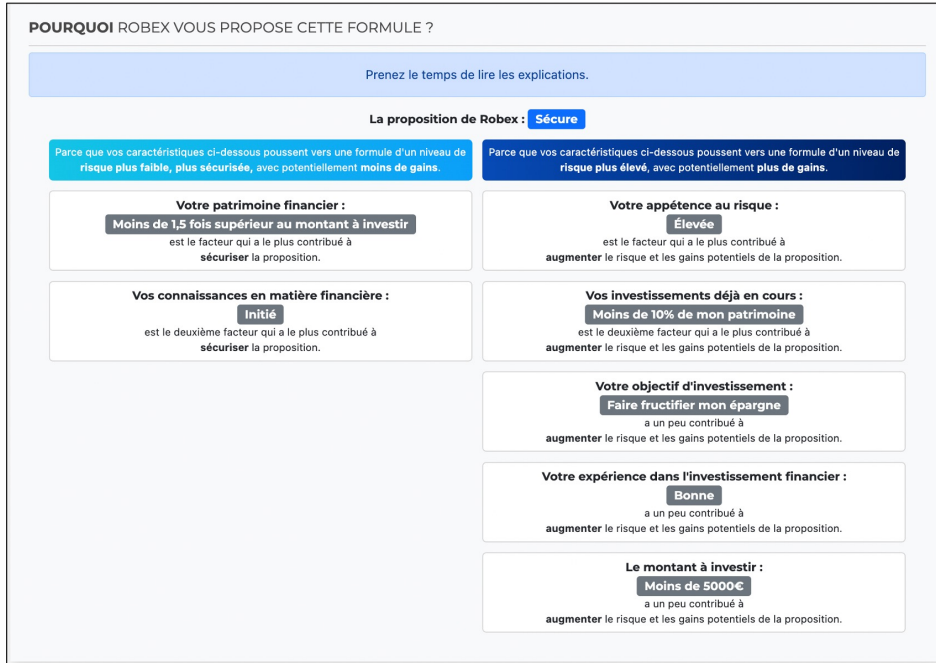


Figure 5.5: Screenshot of the feature-based explanation prototype for Robex. In original language (French). Individual factors that decrease investment risk are shown on the left in descending order of importance and factors increasing investment risk are on the right.

three parts: a semi-structured interview, a task-oriented think aloud portion and a post-study questionnaire. One researcher was present during all interviews and took detailed notes of the participants' answers and think-aloud statements. The first part of the session consisted of a semi-structured interview to explore the needs of life-insurance clients for explanations of recommendations. Structured questions varied slightly if participants were supervisors or novice end-users. Regulatory supervisors were asked about the role of explanations in enabling users to make informed decisions. They were also asked about the best format and type of explanation to achieve this goal. Additionally, they were asked to provide their thoughts on the explanations currently offered by robo-advisors and how to adapt to clients with little financial knowledge. We asked novice users if they had any experience in using robo-advisors or in receiving financial investment recommendations and what explanations they would like to receive about the recommended financial product. We gave some context on life-insurance and on robo-advisors to people that had no experience at all with financial investments. During the second part of the study, participants were asked to use Robex. Participants were observed by the researcher and asked to think aloud throughout their interaction with the system. Finally, participants were asked about their overall impression of the system.

Participants. We conducted interviews with 11 participants: 6 consumer protection experts¹⁵ and 5 end-users.

The consumer protection experts were volunteers from the consumer protection department of the ACPR, the French regulatory authority for banking and insurance services with whom we collaborated for this study. All participants had strong experience in auditing insurance providers (from 3 to more than 10 years). Their expertise and role is to verify that insurance distributors respect "the rules intended to ensure the protection

¹⁵ Four of them were different from the 4 persons we interviewed to design the Robex algorithm.

of the customers" as well as the "adequacy of the means and procedures which they implement for this purpose" and to promote fair commercial practices among industrial professionals¹⁶. Half of them had some experience in supervising robo-advisors.

The novice users were volunteer doctoral students recruited through the network of the university with which the authors are affiliated. All participants received a consent form informing them of the study objectives and identified risks. All participants were volunteers, not compensated, recruited through an email describing the objective and duration of the experiment. An ethics committee was not required for this study.

Inductive content analysis. We conducted an inductive [Elo and Kynäs, 2008] content analysis of the detailed notes taken by one author during the interviews with supervisors and end-users. One author identified concepts and themes about the characteristics of the explanations that emerged from reading the interview notes. First, the author observed that participants talked mainly about either the explanation implementation or the explanation's purpose (notably with discussion around risk). On this basis, different themes for either explanations' format/content or explanations' purpose could be derived that encompass most of the concepts mentioned by participants. The translation from French to English was done after the final categorization.

¹⁶ <https://acpr.banque-france.fr/en/customer-protection/professionals/customer-protection-principles>

5.4 *Study 1 Results*

5.4.1 *Understanding explanation needs from two perspectives*

We grouped the main identified themes of the explanation requirements according to their connection to the format or content of the explanation. Through the supervisor's view, we were able to gather domain perspectives that end users alone would not necessarily have provided, such as understanding the interests of different stakeholders and potential misalignment, where the vulnerability of certain users can be exploited, or the wide range of best practices seen for recommendations and explanations. Conversely, the end-users' perspective reminds us of what clients truly care about, regardless of existing regulations. While the main focus of the supervisors was on the notion of risk, the main concern of the users was not as clear. For some, it was the performance of the proposed contract, for others the reliability of the robo-advisor, and for others, the risk. We discuss below some themes that emerged from both perspectives.

5.4.2 *Redesign principles drawn from the co-design sessions*

Give more precise explanations. The supervisors reported an increasing trend for automated online robo-advisors, and a lack of "good" automated explanations to support those tools. Current robo-advisors' explanations were seen as very "generic" and "nebulous" in general. One of the reasons is the use by many brokers of a third-party software to produce explanations and recommendations, over which they have little control. Supervisors also reported the difficulty for brokers to produce explanations with the increasing complexity of their tools: "There's too much complexity even for them." This highlights the relevance of the XAI domain to help solve real-world problems, even when the underlying recommendation system is AI but rule-based.

Inform customers of the risk. The supervisors insisted on the importance of explanations as a safeguard to inform customers about risk, taking as an example cases of overestimation of the risk for vulnerable people. Supervisors used to phrase "prise décision éclairée", which can be translated literally into English as "enlightened decision-making", to describe the aim of the explanations. This French phrase conveys a stronger notion of user empowerment than "informed choice".

Rule-based algorithm improvement. The supervisors we interviewed also gave us feedback on the rule-based algorithm that we developed. During algorithm testing, they deliberately simulated specific vulnerable user profiles to verify that Robex's recommendation was low-risk. This enable us to add several exceptions to our rules such as if users' risk appetite is 0/7, redirect the user to the most secure proposal regardless

Explanation aspect	Supervisor view	End-user view
Format		
Schematic	"schematic", "graphics and diagrams [for novice users]", "playful", "step-by-step"	"I want to see the scale of the risk, and where I'm placed on that scale"
Content		
Synthetic <i>vs.</i> exhaustive	short, simple, readable, "[Explanations] are a sort of synthesis", "clean and clear" <i>vs.</i> exhaustive, "Just putting a sentence " <i>considering this and that...</i> " is not enough", give links to more information, give enough documentation	simple, "Something that tells you "this is really the points you need to know""
Adapted vocabulary	"adapt vocabulary", "not too much text", "avoid financial jargon"	"use simplified language, not the language of a banker", "need to have more familiar language", "I'm not sure what a placement is"
Purpose		
Justify	link user characteristics and product, "justification", "real need of transparency" motivated by misalignment of interest between insurers and clients, prevent "scams", "what it is based on?"	"Why are you making this recommendation? What factors are you basing it on?", "I want an explanation only if there is a disagreement."
Warn	control, notify, warn, inform, "tendency to underestimate [the risk]", "Explanations are useful because there is a risk.", "the [human] advisor will not say everything", "robo-advisors don't have enough safeguards", "make them [the users] understand that there is a step to take, make them question "do I agree?""	"What are the risks?", "How much do I concretely risk losing on the 50,000 I put in?", "What can I expect in terms of risks and benefits?"
Engage users		"It looks boring", "I'll open them [the links] and probably not look at them."
Teach	enable users to have answers to their follow-up questions	"I don't know anything about that.", "I neither agree nor disagree because I don't really understand this financial concept", "I don't understand this field"

of the other parameters, and if the objective is to protect my savings, cap the recommendation at the second safest.

Support user engagement and learning. Although we could group both supervisor and end-user perspectives into common themes, some themes were discussed more by one group. For example, end-users expressed their need to be engaged—some felt either overwhelmed or bored by the topic. supervisors talked about the need for complete information although end-users insisted on their need for simple, easy-to-digest information, that used simple vocabulary. One participant said that he found it difficult to understand what the numbers or ranged used in the explanations represented because he had no concept of scale in this area. For example, it was difficult to make sense of "less than 30% of my assets". Is that a small, a large portion? This makes it difficult to assess explanations.

Table 5.2: Main themes emerging from the content analysis of supervisors and end-users interviews, with corresponding lexical field and citations.

Find the balance between text and graphics. One of the themes we found was the need for schematic explanations on the one hand and the need for more human explanations that can answer a wide range of users' questions on the other. Two supervisors very much appreciated our graphical, Shapley-based explanations, finding they had never seen something like that in the market and that it responded well to the need to link users' characteristics to the recommended product. However, many—supervisors and end-users alike—indicated their need to be able to chat with a human counsellor despite the explanation. A supervisor also imagined explanations could look more like a Frequently Asked Questions menu and a participant said "I can imagine a chatbot with someone behind it who can answer my questions." This led us to try to balance between text and graphics, following Szymanski et al. [2021]'s findings, and to compare more "conversational" or more "graphical" explanations in the next study.

Clarify visually and accurately the feature's impact. Some participants commented that it was quite difficult to understand what the two columns represented. They would have liked more visual clues, with arrows as in the original Shap explanation, to indicate the direction of each feature's effect. One participant also expressed that she would trust an explanation that correctly scaled the effects of each impact.

Redesign specifications. Based on the legal requirements for explanations and the analysis of supervisors' and end-users' expressed needs, we derived the following elements for the redesign of our explanations.

- *Risk of the recommendation.* We added the risk score of each user from the rule-based algorithm of Robex, and reported it on a scale of one to five to make it correspond to the five recommendations. We added the user risk score and risk scale below the visualisations of the five recommendations.
- *Important Definitions.* As highlighted by end-users and supervisors in Study 1, and by prior work [Bove et al., 2022], it is essential to give the minimal background knowledge necessary to understand the financial concepts used in the recommendations and explanations. We therefore provided on-demand definitions for all important financial concepts through information buttons.
- *Vocabulary.* As pointed out by a non-expert participant, we simplified the vocabulary used in the text. Initially, it contained some financial jargon that we had learned from our informal talks with regulatory supervisors.
- *Descriptions of the effect of complex user input parameters.* Robex used five user input parameters: "Your risk appetite", "Your level of financial knowledge", "the amount to invest proportionally to your total financial assets", "Your financial objective" and "The portion of your financial assets already invested". Out of those five parameters, we saw in Study 1 that the last three were more complex to interpret. For

each of these concepts, we provided (1) the effect it should have on the proposition—either lower or increase the risk the customer can take—(2) an indication of the magnitude of the user's input (e.g. "75% is a very big portion"). An example is shown in Figure 5.6.

- *Direction and scale of the impact of features.* We have converted our original tabular visualisation into a tornado plot to make the direction and scale of the features' impact clearer.

Study 2 tests two additional formats to explore the optimal balance between text and graphics: an interactive graphical format and a chatbot-style dialogic format with a few graphical cues.

5.5 *Study 2 Methodology: A deception-based between-subjects experiment*

In this study, we expand upon the results of Study 1 to examine the usefulness of legally required feature-based explanations in the context of life insurance to help lay users appropriately rely on robo-advisor recommendations. Specifically, we conduct a between-subjects experiment with deception to test for overtrust and overreliance effects. Below we describe the design of the quantitative study, explaining the rationale for the use of a 2x4 factorial design and a between subject crowd-sourced survey.

5.5.1 *A 2x4 factorial design*

Experimental conditions. We used the results of Study 1 to refine the original Robex explanation prototypes and to create different explanation conditions for comparative evaluation. Study 1 led us to question the right balance between text and graphics. Additionally, we build on the findings of the explainability literature presented in Chapter 4, according to which interactivity improves the usefulness of explanations. Specifically, we want to test two types of interaction identified in our taxonomy: "simulate" and "ask" interactions, which have not been directly compared in the existing literature. Therefore, our explanation conditions vary in terms of interactivity and balance between visuals and text. In this quantitative analysis, we examined four distinct explanation conditions.

1. *Control.* Some participants did not receive any explanation. They served as our control condition.
2. *Graphical-static.* The "graphical" explanation we had initially prototyped for Study 1 was improved based on participants' feedback and the redesign specifications outlined in Section 5.3.
3. *Graphical-mutable.* We implemented a version of the graphical explanation where user could change a few parameters that were actionable such as investment amount, objective and portion of assets invested elsewhere. This interaction corresponds to the "mutate / simulate" interaction described in our interaction taxonomy in Chapter 4.
4. *Dialogic.* As somme end users and supervisors compared Robex's explanations to those of a human advisor, we also designed more human-like explanations, i.e. "dialogic" ones. This approach has been adopted in previous XAI work by [Hernandez-Bocanegra and Ziegler, 2021, Hepenstal et al., 2021] for "conversational" explanations. It corresponds to the "ask" interaction in our taxonomy. The dialogues were not responses to free text input from the user, but responses to predefined questions. The user would first see the list of these predefined questions formatted like individual SMS text on the user side of the conversation (in blue) and could click on any of these questions to see

the answer on the Robex side of the conversation (in grey). The answers to each predefined question was also predefined but adapted to the user's characteristics and recommendations. After having clicked on a question, the user could click on any of the remaining predefined questions.

Participants were divided into four groups corresponding to these four different interfaces. The same contextual information was delivered across all the different explanation conditions.

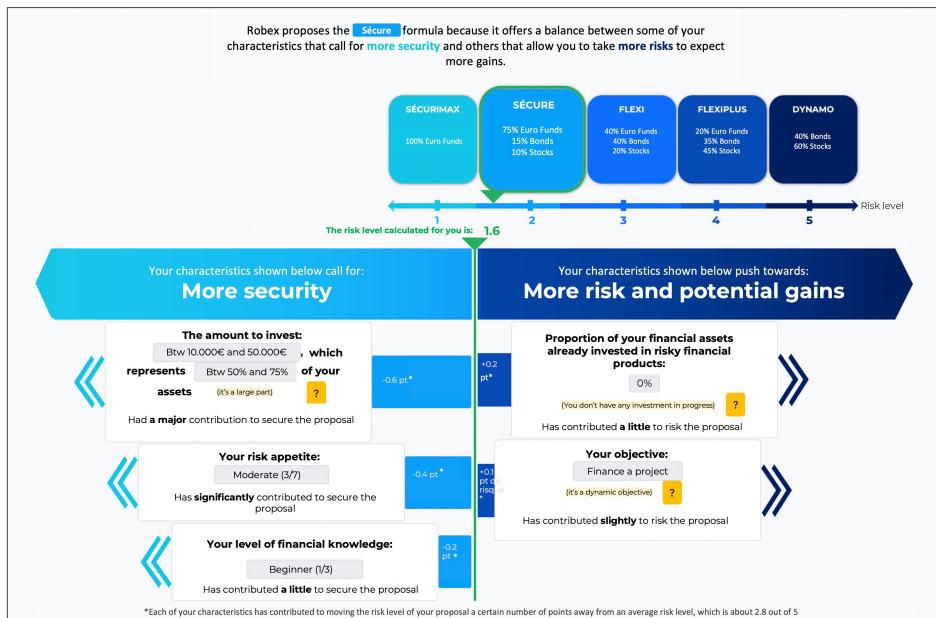


Figure 5.6: Explanation interfaces for each of the condition A "Graphical-static": users see a graphical summary of how their characteristics impact the risk of the proposal. Translated from French to English.

Additionally, as we wanted to test for overreliance and overtrust, we introduced deceptive recommendations as an experimental condition. The objective was to compare the ability of users of different interfaces to detect a crude recommendation error. Each of the four explanation groups described above was divided in two:

1. *Reliable recommendation.* One group received a correct recommendation. These were delivered through the building of a rule-based
2. *Deceptive recommendation.* The other group a false recommendation. The false recommendation was produced by altering the score-based algorithm so that the recommendation was either much too risky or really not risky enough. This was done by altering the initial user's risk score calculated by Robex by a roughly 50% change. The direction of the change was so that more-than average risk-takers were redirected to low-risk proposals and vice versa. For example, if a participant was recommended "Securimax" by the normal Robex algorithm, her risk-score would be increased artificially so as to output the "Flexiplus" recommendation. On the contrary, participants for whom the initial correct recommendation was the more risky "Flexiplus" would be recommended the more conservative "Securimax" product. For participants who initially got the "Flexi" recommendation, if their risk-score

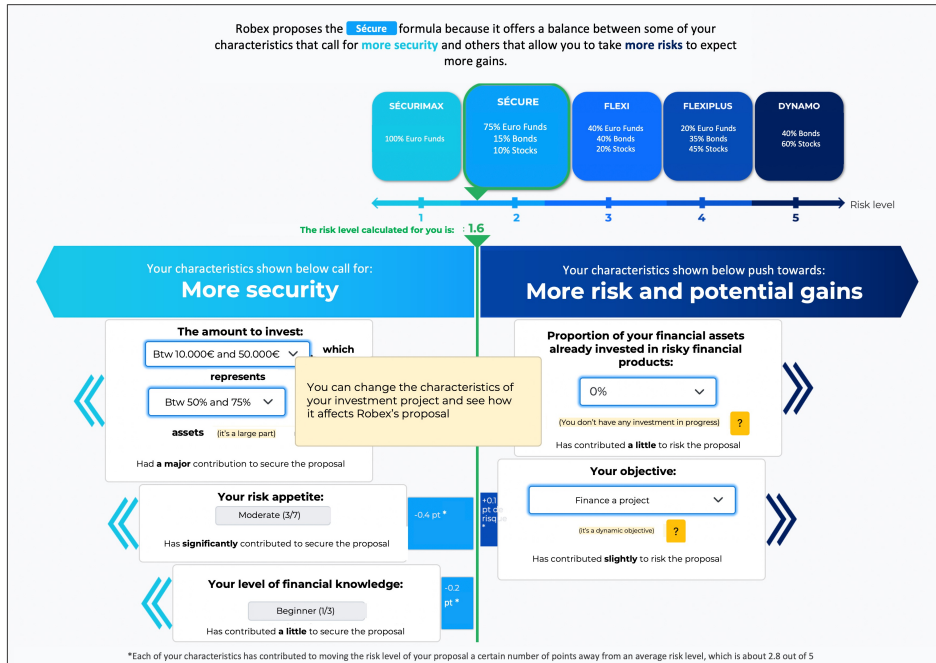


Figure 5.7: Explanation interfaces for each of the condition B "Graphical-mutable": users first see the graphical-static interface and then a pop-up message indicates they can change some of their characteristic. Translated from French to English.

was below 12—out of a maximum score of 21—, they were redirected to "Dynamo" and for risk-scores above 12, to "Securimax". The modified Robex algorithm is presented in the appendix B2.

The explanations of the false recommendation were produced in the same way as the correct recommendations, using agnostic SHAP feature importances based on the skewed Robex algorithm. As a result, the explanations for false recommendations were illogical, such as "Your risk appetite: low (1/7) contributed to increase the risk of the recommendation" *cf.* Figure 5.9.

Measures. Building on prior work conducting empirical studies to evaluate XAI systems [Buçinca et al., 2021, Shin, 2021, Lai et al., 2021, Liu et al., 2021], we measured the concepts described below. We tested the Cronbach's alpha's for the different sets of questions to verify the internal consistency of the questions asked for each dimension. The questions are reported in Table 5.3.

- **Reliance.** Reliance was measured by asking participants if they thought the robo-advisor's recommendation was adapted to their need or not. We were able to measure overreliance when the participant followed an incorrect recommendation.
- **Trust.** Trust was measured through the five question items from the benevolence and competence aspects of McKnight's framework [McKnight et al., 2002]. One item was added to measure if participants felt the need for any additional human advice. overtrust occurred when the participant trusted an incorrect recommendation.
- **Cognitive load.** Cognitive load was measured through the mental demand and effort items of the NASA-TLX Index.


C What would you like to know (click on a question)?

How does **Robex** work?

Robex is an algorithm whose goal is to propose to the user a life insurance plan with an **adapted risk level**. For each user, Robex calculates a risk score and proposes a plan with a corresponding risk level.

Robex recommended to you the **Secure** formula, which corresponds to a risk level of about 2 out of 5.

More precisely, the exact risk level calculated for you (based on your responses to the questionnaire on the previous page) is **1.6**.



The risk level calculated for you is: **1.6**

What are my characteristics that **decreased** the risk of the proposal I received?

Some of your characteristics have contributed to **lower the risk of the proposition you received** (compared to an average risk score of 2.8 out of 5)

The amount that you want to invest represents a large part of your total financial assets: of your assets.

The larger the amount to be invested, the lower the risk of the proposal.

This factor thus had a **major** contribution to **secure the proposal**.

Your risk appetite: has **significantly** contributed to **secure the proposal**.

Your level of financial knowledge: has a **little** contributed to **secure the proposal**.

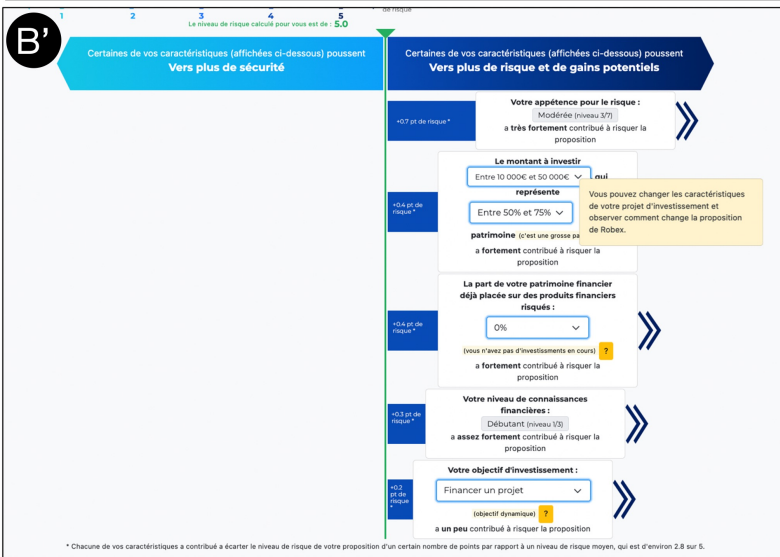
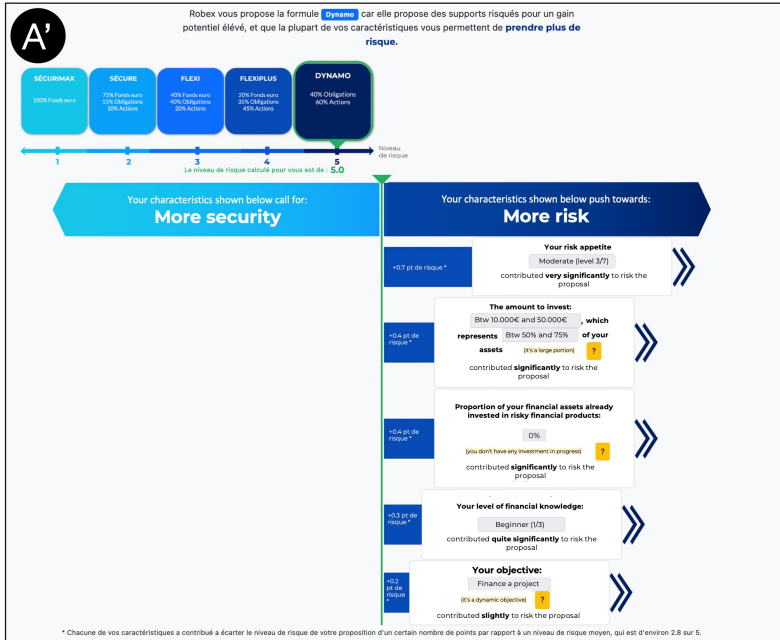
Towards Less Risk <<

The amount to invest compared to your assets Strong impact	Your risk appetite Strong impact	Your level of financial knowledge Small impact
---	-------------------------------------	---

What are my characteristics that **increased** the risk of the proposal I received?

Some of your characteristics have contributed to **increase the risk of the proposition you received** (compared to an average risk score of 2.8 out of 5)

Figure 5.8: Explanation interfaces for each of the condition C "Dialogic": the same information provided in the interfaces A and B) is delivered through "sms-like" textual messages. Some graphics are added to facilitate the visualisation of the risk and of the variables decreasing and increasing the risk of the proposal. Translated from French to English.



- *User engagement.* Three user engagement question items were adapted from O'Brien and Cairns [2015]'s framework. Two items were taken from the Felt Involvement (FI) category and one from the Novelty category (NO).
- *Objective understanding.* Understanding of the recommendation on the one hand and understanding of the explanation on the other were measured through "test" questions. The question about the recommendation was developed by the authors relying on their knowledge of the field and discussions with experts. To measure understanding of the explanation, we used three questions to test if they understood the direction of the impact of some user inputs, as seen in prior XAI work [Szymanski et al., 2021].

Figure 5.9: Explanation interfaces examples for an incorrect recommendation for each of the three conditions: A' "Graphical-static"; B' "Graphical-mutable"; C' "Dialogic". The correct user profile in this case would have been "Secure", but the skewed Robex algorithm outputs "Dynamo". Only A' is translated from French to English, the rest are in original language.

Measure	Questions with [possible responses]	Cronbach's alpha
<i>Understanding of recommendation</i>	What is your estimate of the euro fund percentage in the proposal that was made to you? [Several proposals]	NA
	On a scale of 1 to 5 (5 being the most risky), how risky do you think the Robex proposal is?	
	What is special about a euro fund? [it offers a high expectation of gains for a high risk of loss, it is mostly composed of actions, it is guaranteed by the insurer, I do not know]	
<i>Understanding of explanation</i>	Of your characteristics and goals, which factor weighed the most in the proposal the algorithm offered you? [Several proposals]	NA
	How did the proportion of your financial assets already invested in risky financial products, which is for you ... , impacted the risk of proposal made by Robex? [Increase / decrease / neutral]	
	How did your investment objective, which is ... impacted the risk of the proposal made by Robex?	
<i>Trust-Benevolence</i>	I think Robex is acting in my best interest	0.854
	Robex wants to understand my needs and preferences	
<i>Trust-Competence</i>	Robex is skilled and effective in providing life insurance recommendations	0.878
	Robex has the expertise to understand my needs and preferences	
	Robex is fulfilling its role as a life insurance advisor very well	
<i>Trust-Other (not used)</i>	I would need a human advisor to help me choose a life insurance plan	Not used
<i>User engagement</i>	I felt involved in my task of choosing a life insurance plan	0.818
	The content of the life insurance recommendation site has attracted my curiosity	
	I was interested in the experience	
<i>Cognitive load</i>	I found it mentally demanding to read and understand the proposed life insurance formula	0.829
	I had to make an effort to read and understand the proposed life insurance formula	

Table 5.3: Question used for measuring different metrics with Cronbach alphas (translated from French to English).

5.5.2 Survey procedure and analysis

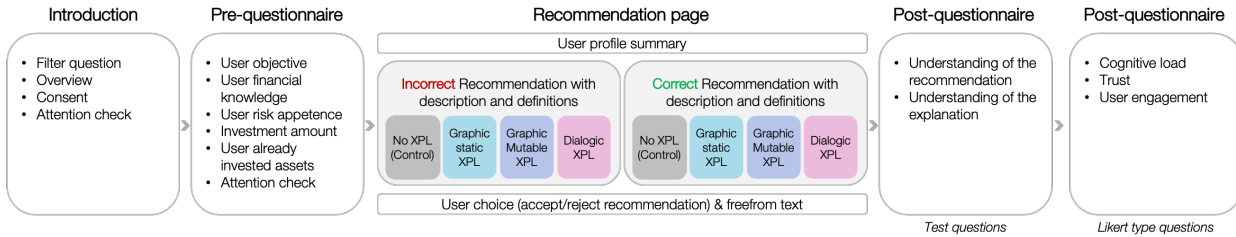


Figure 5.10: The workflow of our quantitative experiments. The profiling questionnaire is used to produce a personalized recommendation of a life-insurance contract. Clients can review the recommendation, the explanation and then decide to follow the recommendation or not.

Procedure. Our goal was to target participants who might be life insurance robo-advisor users. As participants were crowd-sourced, we began with a selective question to filter out users who were not likely to be users of life-insurance in the near or distant future. The question used was "To begin with, we would like to know how you feel about life insurance: 1 - I might sign up (for the first time or again) to life insurance in the near or distant future. / 2 - I am not considering signing up (for the first time or again) to life insurance in the near or distant future, even though I'm curious to find out more on the subject." The answers were formulated so that it was not obvious to guess which answer to select to be able to continue. Only participants who checked the first answer were selected to continue. On the crowd-sourcing platform, participants were asked about their highest level of education and gender. Participants were redirected to Robex and provided with an overview of the study. They were asked to provide their consent to participate and then underwent an attention check. The two following steps in the study process replicate what we can see in existing robo-advisors: a profiling questionnaire followed by recommendation page. Participants had to go through the profiling questionnaire. They were then distributed randomly in our eight different conditions as shown in Figure 5.10, which illustrates the experimental workflow. They read through their user profile summary at the top of the page, the description of the recommendation. If applicable, they saw an explanation of why this recommendation was made to them, and then they had to choose whether to accept or reject the proposed life-insurance plan. We also collected their qualitative feedback about explanations through a short free-text field. Finally, a two-page post-questionnaire measured their understanding, workload, trust and engagement in using Robex.

The whole study lasted around 10 minutes. Participants were paid around 3€50¹⁷ for completing the study. We randomly assigned participants to an experimental condition until we had reached a minimum of 30 participants in each of our eight conditions.

Participants who failed attention checks, took less than 5 minutes or wrote non-serious content (repeated keyboard strokes, clearly ironical or

¹⁷ Lucid goes through several suppliers to gather participants. Each supplier receives 3.50€ for each study completed, takes a commission and pays the rest to the participant.

insulting content) in the free-text field were excluded. We also implemented time counters: participants could not continue to next page if a (small) minimum amount of time had not elapsed. In addition, on the recommendation page, we set time counters for each of the three sections of the page: profile summary, recommendation and explanation. The time thresholds were calibrated to correspond to a quick reading of each section. After the time had elapsed, a button appeared to say "OK continue" or "Show recommendation". These time counters therefore also served as a way to gradually disclose content and avoid cognitive overload [Springer and Whittaker, 2019]. This was to make sure that participants read through the profiling questionnaire, the recommendation and the explanation. We ended up with 32 participants in each condition.

At the end of the survey, participants in the deceptive condition were informed that they had received a wrong recommendation. All participants were reminded that the financial advice presented was fictitious and non-relevant for their personal needs. The study was approved by an academic research ethics committee.

Participants. French workers between 18 and 65 years old were recruited online through the platform Lucid¹⁸. Of the study respondents that were finally included in the survey, 73% were female and 27% male—although some participants did not provide any answer to that question. 61% had an undergraduate or a graduate degree (Bachelor, Master, Doctorate and other specialized education). We cannot explain the skew towards women participants but it is possible that more male participants did not want to answer this demographic question or that our filters about the interest in life-insurance or seriousness of the responses excluded more male participants. Participants had an average financial knowledge score of 1.3 out of 5, and were therefore for the most part representative of non-expert users. Financial knowledge was measured in the pre-questionnaire through specific questions written with the help of four supervisors from the French Regulation Authority of financial services (*cf.* Table 5.1 for the detail of the questions).

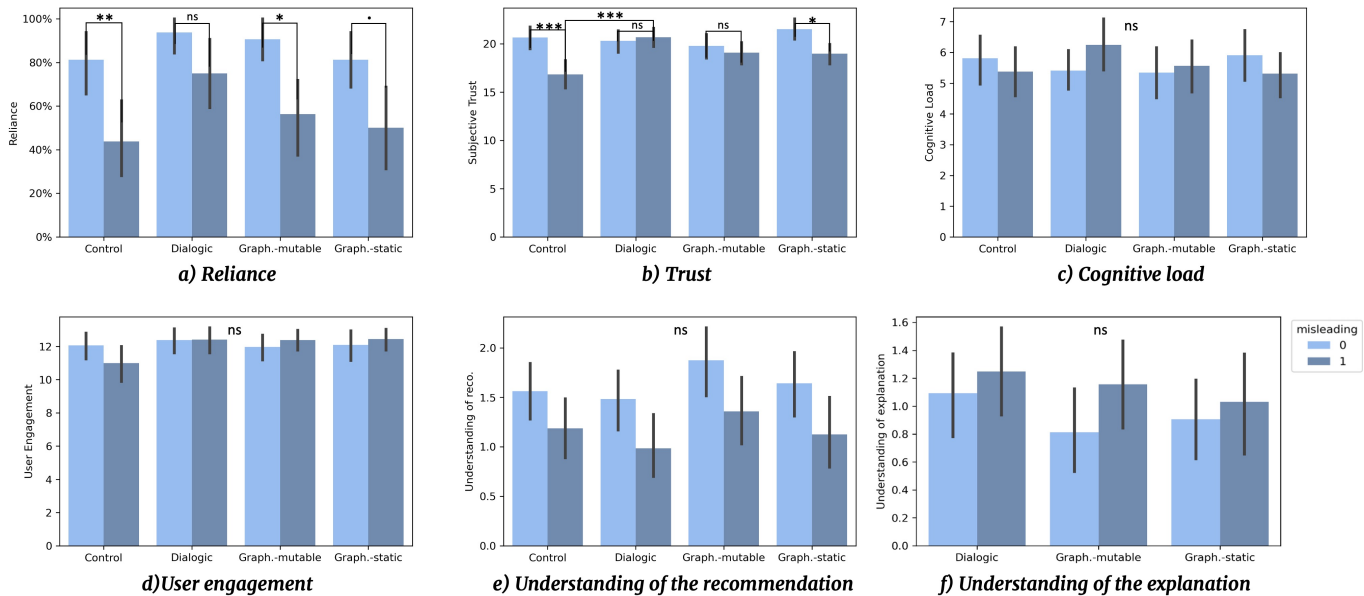
¹⁸ <https://lucid.co/>

Analysis. For all evaluation measures, we ran a two-way ANOVA analysis with the explanation conditions and the recommendation conditions (correct or false) as the independent variables. Our eight groups had a minimum of 32 participants in order to confidently meet sample size considerations for ANOVA. For groups that had more participants, we randomly selected 32 responses. When significant, we conducted post-hoc Tukey's HSD test for pairwise comparisons. We used the Shapiro-Wilk test to check that the assumptions for ANOVA were met and the Bartlett test to verify the homogeneity of variances. We also controlled for socio-demographic confounding factors: education, age, and gender as control variables, although this data variables was incomplete.

5.6 Study 2 Results

All Cronbach's alphas for the different sets of questions were significant, except for trust: we had to remove the question about the human advisor (we initially thought this question could be related to trust in the robo-advisor, as it measured trust in a (human) advisor, but it was a false intuition). For all the evaluation measures, the residuals of the regression showed a near-normal distribution, as confirmed by the Shapiro-Wilk test, validating the assumptions for ANOVA. Additionally, the Bartlett test indicated that variances were homogeneous.

Figure 5.11: Results for Study 2. Vertical lines represent the 95% confidence interval. Asterisks and dots indicate the statistical significance of the results: *** $p\text{-value} \leq 0.001$, ** $p\text{-value} \leq 0.01$, * $p\text{-value} \leq 0.05$, • $p\text{-value} \leq 0.07$, "ns" non significant.



5.6.1 Explanations do not help to better calibrate trust

We found that the no-explanation control group was more or equally likely to distinguish between good and bad advice than the explanation groups. We found a statistically significant difference in trust ($p=0.001$) and reliance ($p=0.01$) between the no-explanation control group that received a correct proposal and the no-explanation control group that received an incorrect one. However, we did not always observe this with participants who received explanations. Specifically, there was no statistical difference in trust and reliance on the advice between the dialogic explanation group that received a correct recommendation and the dialogic explanation group that received an incorrect recommendation. For the graphic-mutable explanation, we found participants were able to calibrate their reliance on the advice between the incorrect and correct proposal ($p=0.03$), but not their trust. In the graphic-static explanation condition, people trusted a correct proposition significantly more than an incorrect one ($p\text{-value}=0.05$) and relied on the correct proposition almost but not significantly more ($p=0.064$) than on the incorrect one. Overall,

out of those three explanations, it may be the graph-mutable explanation that performed best, because it enabled participants to appropriately calibrate their demonstrated trust, *i.e.* reliance on the recommendation. However, none of the explanations outperformed the control condition in appropriately calibrating trust and reliance.

5.6.2 *Dialogic explanations increase subjective trust*

We found that users who were shown an incorrect recommendation and a dialogic explanation trusted significantly more the robo-advice compared to the no-explanation group ($p=0.001$). Further, we found that participants in the incorrect recommendation and dialogic explanation condition were almost significantly ($p=0.068$) more likely to rely on the incorrect robo-advice than participants in the incorrect/control condition.

5.6.3 *Dialogic or graphical explanations do not improve user understanding*

The different explanation formats did not improve users' understanding of the recommendation and more specifically its risk—question 1 out of 3 measuring recommendation understanding (cf. Table 5.3). Based on the graphs in Figure 5.11, there appears to be a tendency for graphical-mutable explanations to lead to better understanding of the recommendation than other conditions, but the effect was not significant ($p=0.1$). Further, the level of understanding of the explanations was comparable across the different explanation conditions. However, people in the deceptive conditions were significantly less likely to understand the characteristics of the recommendation and the explanations ($p=0.001$). This result is based on one-way ANOVA with solely the recommendation condition (correct or false) as the independent variable.

This evidences that people are less likely to understand a recommendation that is not suited to their needs, or that they did not expect.

5.6.4 *Explanations do not affect cognitive load and user engagement*

We do not find any statistically significant effect for the different explanation conditions on users' subjective cognitive load and user engagement. This finding contradicts other work on the cognitive cost of explanation [Vasconcelos et al., 2022]. Perhaps this is the case here because understanding financial recommendations is already cognitively demanding enough due to the complexity of the field, and the cost of adding explanations is negligible in comparison—average perceived cognitive workload for using the robo-advisor was 5.6 out of 10. The "simulate" and "ask" interactions did not improve users' subjective engagement in the task. This may also be explained by the seriousness and unamusing nature of this specific task in the finance domain.

5.6.5 Higher levels of education reduce overreliance

As shown in Figure 5.12, we conducted an additional analysis to study the effects of education. Indeed, while controlling for confounding factors, we had noticed that education could play a role in trust and understanding of the recommendation. The original categorical data collected on the education crowdsourcing platform included eight different categories representing levels of education in French. 68 participants out of 256 respondents did not provide their education levels. We created larger groups by combining educational levels equal to or less than the "Baccalauréat", which is the equivalent of a high school diploma in France. Educational levels one to three years after the Baccalauréat were grouped together. Masters and doctorates, corresponding to more than four years of education after the Baccalauréat, formed a third group. To run a two-way ANOVA with education (three groups) and recommendation conditions (two groups) as independent variables, we checked the minimum number of participants in these six groups m and randomly selected participants in the largest groups to form six groups of size $m = 20$. The results from the two-way ANOVA indicated that participants in the highest education level group tended to rely significantly less on the wrong recommendation compared to the correct one ($p=0.05$). This indicates that education plays a role in critical thinking and ability to exhibit a healthy dose of skepticism. In addition, we found that participants in the lowest educational group understood the incorrect recommendation significantly less than the correct one ($p=0.03$).

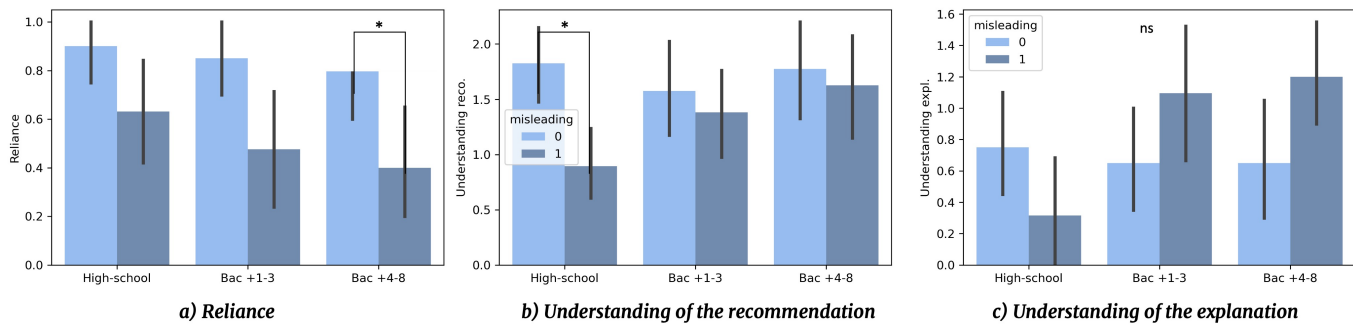


Figure 5.12: Effects of education on reliance, understanding of the recommendation and of the explanation.

5.7 Discussion

5.7.1 Dialogic vs. Graphical explanations

According to Miller [2019], explanations are best provided through a social process, *i.e.* a conversation, because it matches the way humans explain things. In fact, "dialogic" explanations have been favorably presented in the XAI literature. For example, Hernandez-Bocanegra and Ziegler [2021] presented how dialogic management systems can respond to users' questions about a hotel recommender system, and Hepenstal et al. [2021] showed how conversational explanations can be useful for criminal investigators. While the benefits of dialogic explanations might be real regarding user satisfaction and explanation usefulness in some contexts [Hernandez-Bocanegra and Ziegler, 2021, Hepenstal et al., 2021], our results, in turn, shed light on the overtrust downside of "dialogic" explanations for clients of online recommender systems. It is possible that either the "humanness" of the dialogic explanation we presented, or the familiarity of users with chats, made them more inclined to accept robo-advice. In fact, some people might see the anthropomorphisation of systems as suspicious. One of our end-user participants in the pilot Study said that "*It's quite a lot of anthropomorphization*". This is consistent with the study by Hepenstal et al. [2021] in which participants were uncomfortable with the humanness of the XAI agent and wanted to have it clear that they were not talking to a real person. Our findings also qualify Szymanski et al. [2021]'s results according to which participants prefer graphical explanations but understand textual explanations better. The authors further advance that hybrid textual and graphical formats could improve both user satisfaction and understanding. Our study qualifies this result by showing that users made less mistakes with graphical formats which presented small amounts of text than with dialogic formats with small amounts of graphical visualizations. This contrasts with Szymanski et al. [2021]'s finding that text is better understood—however the textual explanations in this work were much shorter. Perhaps the brevity and the synthetic aspect of our graphic explanations compared to the dialogic explanations were instrumental in improving users' appropriate reliance.

5.7.2 Legal requirements for feature-based explanations

In this study, we showed how legal requirements to justify investment advice based on client's features may take shape using a classical XAI method (SHAP) and various explanation representations. We further found that the legal sub-objectives of the explanation that we defined in Section 5.2.3 to help users make "fully informed" decisions were not fully achieved. Users were not better able to 1) appropriately rely on the recommendation, 2) understand the recommendation or 3) appropriately calibrate their trust in the robo-advisor compared to the control condition.

As noted in Section 5.2.3, the objective of the law requiring insurance intermediaries to specify in writing "the reasons for the appropriateness of the proposed contract" is also to discipline brokers by making non-objective, self-interested, recommendations more visible and punishable. Feature-based explanations are therefore not useless, because they at least serve the purpose of disciplining insurance intermediaries by forcing them to show how the proposed product corresponds to the customer's risk profile.

However, our work changes the perspective on the benefit of explanations for customers' understanding and reliance. Explanations are not always "all good", they must be designed so that overtrust and overreliance effects are mitigated. If the explanation formats we presented could not meet the legal objectives we highlighted, future work could address how to design explanations that are cognitively engaging for lay-users. We develop this in Section 7.2 of Chapter 7.

5.8 *Limitations*

This work has some limitations. First, the content analysis in Study 1 was performed based on the detailed notes that one author took during the interviews, which may have limited the amount and breadth of captured input from participants. In addition, the non-expert participants from the qualitative study were graduate students, who represent a very specific sample of non-expert users. One of the limitations in our domain-driven contextual enquiry is that we used a simplified and fictional life-insurance robo-advisor. Some factors such as time horizon, detailed descriptions of the funds, of their historical performances and the costs of each contract were not taken into account. We did this to simplify the building of the tool, and also because we felt adding costs and performances might have diverted participants' focus from the risk of the proposals, which is the most critical information for users to understand according to supervisors and the spirit of the legislation. Future work could explore similar research questions with a real robo-advisor. Additionally, one of the main limitations of crowd-sourcing participants in Study 2 is that they might lack the mental engagement or involvement with the subject. To increase participant engagement, we let them answer the survey with their own profile, instead of presenting a predefined profile for all participants. We verified that the type of recommendation did not have a significant impact on our measures. Additionally, we implemented a question to filter out users completely uninterested in life-insurance, attention checks, text fields and time counters to filter out non-serious participants. Nevertheless, it is possible that the participants in our study were not representative of a real user of a real life-insurance robo-advisor. Also, the participants in our study were also mainly women (73%).

5.9 Conclusion

In this chapter, we carried out a co-design experiment aimed at understanding the needs and requirements for explanations in robo-advisors from the perspectives of non-expert end-users and supervisors in customer protection. Based on these findings, we designed various prototypes of feature-based explanations for online recommendations in life-insurance, including both interactive and static options. We then presented the results of a 2x4 between-subjects experiment to investigate whether different formats of feature-based explanations help novice users to appropriately rely on, trust, and understand life insurance plan recommendations. We found that providing feature-based explanations did not significantly improve users' understanding of the recommendation, or lead to more accurate reliance on the recommendations compared to having no explanation at all. We also found that explanations provided in a dialogic format, where users can choose a question and get chatbot-like text answers, increased users' trust in the robo-advisor and did not significantly improve user understanding. This led us to conclude that graphical formats could be better suited to inform clients. This leaves us in a quite unsatisfactory state of affairs where the obligation to inform clients does not fulfill its promises to empower users in better understanding the recommendation or in making better decisions. However, in regulated contexts such as life insurance, regulators and internal compliance systems act as barriers to the manipulation of user trust, ahead of the protection provided by user self-regulation. The ability to detect untrustworthy recommendations does not primarily rest on the shoulders of end-customers.

In the next chapter, we investigate the explanation needs of financial regulatory supervisors to control the trustworthiness of AI systems.

Chapter 6

Understanding the supervisors' needs for explainable AI in financial crime detection

REGULATORY SUPERVISORS PLAY a critical role in ensuring the trustworthiness of AI systems and preventing end-customers from having to detect false AI recommendations. Rather than mere explanations, supervisors expect "*justifications*" by regulatees that an AI system or decision complies with a legal standard, rule, or objective¹. However, little is known about the actual needs of supervisors concerning such justifications of AI systems.

In this chapter, we take another case study in finance: anti-money laundering and countering financing terrorism (AML-CFT). We take a dual user-centered and legal approach to describe the explanation needs of regulatory supervisors to verify AI compliance with AML-CFT regulation. We examine a socio-techno-legal supervision system in AML-CFT in France, as an example of AI use in a highly-regulated industry. We draw on 6 workshops with supervisors and bank practitioners to outline the auditing approaches of AML-CFT supervisors. Our findings present the AML obligations that conflict with AI opacity. We then formulate seven needs that supervisors have for model justifiability. Finally, we discuss the role of explanations as reliable evidence on which to base justifications.

We begin by presenting the related literature and the relevant background in AML-CFT in Section 6.2. We then describe our methods and findings in Sections 6.3 and 6.4.

This study was made possible thanks to the collaboration of the ACPR, the French regulatory authority of financial services and the Crédit Agricole, a large French bank. The views expressed in this chapter are exclusively those of the authors and the participants of this study in their personal capacity. They cannot be taken as the views or policies of the ACPR or Crédit Agricole.

¹ Cf. Section 1.1.5 in Chapter 1 for a clarification of the terminology employed and the differences between explanation and justification. As noted by [Hildebrandt, 2019] and [Henin and Le Métayer, 2022], justifications are extrinsic as they refer to norms and regulations.

6.1 *Motivation and research questions*

AI regulation has been rapidly gaining interest due to the advances of generative AI and the emergence of new AI regulations². However, highly regulated industries, such as banking, healthcare, or the military, already have structures in place to deal with technological risks. These domains are characterized by well-established norms, experience in putting principles into practice, a common goal of social welfare, and robust professional accountability mechanisms [Mittelstadt, 2019]. In banking, machine learning adoption is on the rise [Financial Conduct Authority, 2019], with regulators sometimes encouraging industry players to consider AI to improve the efficiency of their systems [Board of Governors of the Federal Reserve System et al., 2018]. However, little new regulatory guidance has been provided to address the specific risks of AI [The Federal Reserve Board of Governors in Washington DC, 2011, Financial Conduct Authority, 2022] and firms call for a more proactive regulation approach [Financial Conduct Authority, 2019, Truby et al., 2020]. Truby et al. [2020] notes an overall lack of guidance on AI use from "typically cautious financial regulators". Overall, clarification is needed on how current regulatory mechanisms address the risks of AI.

² For example, the developments of the AI, Digital Services and Digital Markets Acts in Europe and the Algorithmic Accountability Act in the US this year [European Commission, 2021, European Parliament and Council, 2022, Yvette D. Clarke, 2023]

In this study, we focus on a highly-regulated area, anti-money laundering and countering financing terrorism (AML-CFT). AI applications for AML-CFT, such as unsupervised anomaly detection, have attracted increasing attention from both industry players and academics for their potential to reduce compliance costs and detect new patterns of money laundering that current rule-based systems are not aware of [Gupta et al., 2023, Singh et al., 2018]. In experimental conditions, Weber et al. [2018] has found that these methods can reduce the number of false alerts for money laundering by 20 to 30%. The impact of such technologies is all the more promising as current AML-CFT systems are relatively ineffective [Bertrand et al., 2021]. The United Nations Office on Drugs and Crime estimates that between 2 and 5% of global GDP is laundered each year and less than 1% of these funds are seized or frozen [UNODC, 2011]. Banks have been increasingly touting the use of artificial intelligence (AI), to the extent that AI use for AML-CFT is entering a tipping point. In October 2022, a Dutch court ruling confirmed that the financial institution Bunq could use AI despite reservations from the regulator [Trade and Industry Appeals Tribunal, 2022]. Big tech companies have also begun to provide AI services for AML-CFT systems within banks, such as Google's collaboration with HSBC which resulted in a 60% reduction of false positive alerts and quadrupling the number of true positives [Tokar, 2023].

Kruse et al. [2019] argue that the primary challenge posed by AI algorithms in the finance industry is related to their opacity. As highlighted by Kuiper et al. [2021], AI opacity undermines the ability of financial institutions and regulators to control their systems, thereby posing a risk to financial stability, institutional trust and consumer protection [Kuiper et al., 2021, McWaters and Blake, 2019]. In AML-CFT, concerns of regulators have also focused on the lack of transparency in AI models and on

measuring their added value [Gruppetta, 2017]. Overall, it is undisputed that a certain level of transparency is required for AI models [McCaul, 2022]. However, it is rarely specified to what extent and why AI explanations should be generated in relation to applicable legal requirements. Moreover, few studies have explored the regulator perspective, despite the fact that they are an essential audience of AI explanations.

In this chapter, we focus on AML-CFT supervisors in France, who act as the national public auditors of AML-CFT systems in banks. We strive to understand the supervisors' perspective on AI transparency and justifications, in this case in the highly regulated AML-CFT environment in France. Specifically, we leverage two scenarios of promising AI applications from the AML-CFT literature and conceptual design artifacts of AI justifications and explanations [Gaver and Martin, 2000]. We outline the justification requirements and information needs of supervisors regarding AI systems to help banks better design justifications for AI systems and to help supervisors build relevant explainability and testing solutions for auditing purposes. Grounded in the context of AML-CFT, our study is guided by the following research questions:

RQ1: *What are regulatory supervisors' current auditing practices and socio-techno context? (Section 6.4.1)*

RQ2: *How does AI opacity conflict with compliance requirements and to what extent can justifiability address these tensions? (Section 6.4.2)*

RQ3: *What are the needs of supervisors for justifiability of AI systems? (Section 6.4.3)*

Our study adopts two original approaches. First, the needs and context of regulators, supervisors and auditors is not currently well understood. By exploring their justification needs, we can reduce regulatory uncertainty around the use of AI. Investigating the supervisor perspective will inform how existing accountability mechanisms can be applied to AI technology. Second, in order to fully understand the objectives and needs of supervisors, it is necessary to consider the legal requirements. As such, we conduct a multi-pronged socio-techno-legal study of these users and their context.

6.2 Background

6.2.1 HCI work on eliciting user explainability needs

As presented in Section 2.4.3 in Chapter 2, HCI researchers have often relied on interviews and workshops [Sun et al., 2022, Liao et al., 2023, 2020, Ehsan et al., 2021, Maltbie et al., 2021, Tsai et al., 2021, Kim et al., 2023, Ehsan et al., 2019] to learn about the needs and context of specific user groups and inform the design of explainability systems. Additionally, scenario-based design, [Carroll, 1997], in which participants are engaged in a scenario to elicit their feedback, was used multiple times in explainability [Cirqueira et al., 2020, Sun et al., 2022, Wolf, 2019, Liao et al., 2023]. However, very little work has explored the needs of regulators as a user group [Kuiper et al., 2021], and no work in the HCI field has addressed the elicitation of explainability needs using both a scenario-based and a legal approach, to the best of our knowledge. Our view is that it is particularly relevant to the study of the needs of regulators. For example, Chazette and Schneider [2020] emphasised that the elicitation of explainability needs should also take into account laws and norms, cultural and corporate values, domain aspects, organisational constraints such as time, resources, etc [Maltbie et al., 2021].

6.2.2 Designing AI justifications for compliance

As noted by Hildebrandt [2019], explainability is only a small part of the justifiability equation for AI systems and may obscure the bigger picture. However, the notion of legal justification of AI systems has not received as much traction so far. Explainability has received much more attention. Specifically, "legal explanations", *i.e.* explanations designed to support the legal compliance process, have been examined by XAI researchers [Carvalho et al., 2019, Beaudouin et al., 2020, Dupont et al., 2020]. The requirements of the General Data Protection Regulation (GDPR) [European Parliament and Council, 2016] to provide users with "meaningful information about the logic involved" have received much attention from explainability researchers [Hamon et al., 2022, 2020, Bibal et al., 2021, Confalonieri et al., 2021, Doshi-Velez and Kortz, 2017]. Recent work reviews in detail the legal requirements for explainable AI [Nannini et al., 2023, Bibal et al., 2021, Doshi-Velez and Kortz, 2017, Panigutti et al., 2023b]. Nannini et al. [2023] highlight that regulations are informed by coarse notions of explanations. Nevertheless, Doshi-Velez and Kortz [2017] argue that "legal explanations" are technically feasible, mainly through local explanations and counterfactuals. Bibal et al. [2021] presents four levels of explanations to meet the different types of requirements: explanation of the main features, of all features, of the features involved in a decision, or of the whole model.

However, this interdisciplinary body of work, has not yet adopted a user-centric approach to study the needs of regulators, who are the main end-users of such "legal explanations".

6.2.3 *Auditing AI systems*

Some work has emerged to define AI auditing and its role in relation to traditional audits [Sandvig et al., 2014, Metaxa et al., 2021, Toader, 2019] or to outline audit approaches and principles [Sandvig et al., 2014, Koshiyama et al., 2021, Raji and Buolamwini, 2019, Mökander et al., 2023]. Sandvig et al. [2014] first introduced the notion of algorithm audit, with the application of Internet platforms algorithms in mind. Mökander et al. [2023] summarized the promise of AI auditing in three ideas: it is procedurally regular and transparent, it enables proactivity in addressing AI harms, and it is conducted by independent parties. Koshiyama et al. [2021] give four main verticals of algorithm auditing: performance and robustness, bias and discrimination, explainability, and privacy. The first vertical encompasses concepts such as resilience to attacks, fallback plan, accuracy, reliability, and reproducibility. They define seven levels of explainability, corresponding to increasing levels of access to information up to the complete "white-box" setup. Raji et al. [2020] drew lessons for AI auditing from industries including finance. The authors discuss the historical role of internal audits in this domain, and their focus on organisational aspects and risks. They also consider financial auditing to be "lagging behind the process of technology-enabled financialisation of markets and firms". The literature on AI auditing is still in its infancy [Falco et al., 2021], and has so far only focused on definitions and methodological aspects of audits, from a theoretical point of view. Very little research has offered qualitative empirical insights on the socio-techno-legal aspects of AI audits.

6.2.4 *The AML-CFT context*

Overview. Money laundering is the action of concealing the origin of funds illegally obtained. Terrorist financing is a different process: it involves concealing the destination of funds by raising, storing, moving, and using the money [Levi and Reuter, 2006]. To detect these financial crimes, AML-CFT laws require banks to carefully control with whom they are engaging in a business relationship and to actively monitor their customers' transactions [Bertrand et al., 2021]. This implies that banks map out the money laundering risks to which they are exposed, taking into account their activities and customers, and putting in place a detection system, including an often automated "transaction monitoring system" that flags unusual activities. In general, this rule-based approach begins with an alert is first triggered from an automated system usually based on rules (such as "transaction is superior to a certain amount"), then it is quickly reviewed by a human analyst and either closed or passed on to a second level of review. If the alert is still considered suspicious at this stage, a case is created and a more extensive investigation is opened to be reviewed by more experienced analysts. If the suspicion is confirmed, it is reported to the national financial investigative body—TRACFIN in France—which conducts a deeper investigation [Jullum et al., 2020]. If there is evidence of a financial offence, the case is passed on to the law enforcement authorities ³.

³ c.f. Figure 1 in [Kute et al., 2021].

Legal requirements. AML-CFT laws propose a risk-based approach, meaning that banks have to identify the risks they are exposed to and take appropriate measures to mitigate them [Financial Action Task Force, 2007]. The risk-based approach to AML-CFT is widely adopted and has been recommended by the Financial Action Task Force (FATF), the intergovernmental organization dedicated to combating money laundering and the financing of terrorism, to its 39 members, which includes 24 non-EU countries [Financial Action Task Force, 2014]. It is also the standard approach in Europe as it has been recommended by the European Banking Authority [European Banking Authority, 2016].

The banking sector also has "internal control" obligations that constitute a set of safeguards enabling financial institutions to control the risks of their activities [Raji et al., 2020, Soh and Martinov-Bennie, 2011]. EU countries are subject to such requirements under Directive 2013/36/EU. Under these requirements, banks have to implement three "lines of defense" to ensure that their financial activities remain legal: level one corresponds to the day-to-day business operators; level two requires a separate unit responsible for monitoring level one; level three is an audit team that intervenes periodically. If banks fail to comply with these obligations, they can face heavy fines by the national supervisory authority. In France, these fines amounted to several million euros between 2016-2021, sometimes amounting up to 6.5% of the fined banks' revenues [Conseil d'Orientation pour la lutte contre le blanchiment et le financement du terrorisme, 2023].

The role of supervisors. Supervisors are agents of regulation. In France, their role is laid down in the regulation⁴, and described on the French Regulator's website⁵. Supervisors monitor the compliance of financial institutions with European and national AML-CFT laws. They also influence the development of AML-CFT frameworks by synthesizing gaps, threats, and best practices at the national level. For example, the French supervisor annually reports on the threat posed by money laundering and terrorist financing and often publishes guidelines and thematic reviews detailing the supervisor's expectations and interpretations of the law.

⁴In Articles L561-36 to L561-44 of the French Monetary Code.

⁵<https://acpr.banque-france.fr/controler/lotte-contre-le-blanchiment-des-capitaux-et-le-financement-du-terrorisme/presentation-du-controle-lcb-ft>

AI for AML-CFT. Banks have only recently begun to explore the use of machine learning in AML-CFT, but it is one of the most impactful applications of AI in banking [Fritz-Morgenthal et al., 2022]. AI development is mainly due to two factors. Firstly, AI promises better performance than traditional detection systems, which are based on known scenarios of money-laundering schemes. The most promising use is through unsupervised and reinforced learning that have the potential to detect anomalies which shed light on typologies of money laundering that have not been previously reported [Canhoto, 2020]. AI can also help set smarter alert thresholds, help human analysts prioritize alert treatment, and enhance the quality and diversity of the data used in criminal investigations [Chen et al., 2018, Kurshan and Shen, 2021, Labib et al., 2020, Lorenz et al., 2020, Ngai et al., 2011]. Secondly, AI enables banks to cut costs by alleviating repetitive tasks and reducing the human staff required to review alerts

[Overrein, 2020, Singh et al., 2018].

However, AI is still a relatively recent topic in AML-CFT, and AI-based systems have been subject to few, if any, regulatory audits to date. So far, only a handful of national supervisory authorities have expressed positions on AI. In 2018, the Monetary Authority of Singapore stated to be "in agreement that such advanced technologies can and should be leveraged by banks" [Singh et al., 2018]. A report on AI for AML in Norway, however, argues that banks "as well as regulators have historically been reluctant to use AI" [Overrein, 2020]. The Dutch Central Bank (DNB), in November 2022, was hesitant over machine learning technologies for AML as illustrated in a regulatory sanction [Blakey, 2022] but has since cautiously opened the door for its use [Singh et al., 2018, Hoegen et al., 2023]. The French supervisor has not yet expressed clear guidance on AI but has been generally open to the technology. They have also developed an internal AI-based tool to challenge the performance of banks' systems [Laporte, 2021].

Explainability and transparency in AML-CFT. Explainability (XAI) has often been presented as a requirement to meet compliance standards in AML-CFT [Bellomarini et al., 2020, Fritz-Morgenthal et al., 2022, Gerlings and Constantiou, 2022, Al-Shabandar et al., 2019]. In her 2022 speech about technologies to fight financial crime, Elizabeth McCaul, member of the Supervisory Board of the European Central Bank (ECB), presented explainability and transparency as "two of the most important challenges for AI" [McCaul, 2022]. However, the specific requirements for explainability and transparency remain vague and general. It is not yet clear which precise legal requirements they would fulfill.

Nevertheless, several efforts to build explainability solutions have emerged in AML-CFT over the past few years. According to Kute et al. [2021]'s review of AI solutions in AML-CFT, 51% of the scientific papers that present a machine learning method for AML also consider the explainability of their solution, such as knowledge-graphs rule-based reasoning approaches [Bellomarini et al., 2020]. Weber et al. [2023] identify case studies from the literature where AI and XAI were successfully applied in real financial contexts. The paper also stresses that XAI in AML is under-explored. However, the majority of these contributions are in computer science and do not consider the complex realities of the AML-CFT context.

Some studies have provided more detail on users' needs for explainability in AML-CFT. Recent work has emphasized the need to understand why an AI model raised an alert, and understand the main features that drove the decision, for the banks' investigators and the national financial investigative bodies [Al-Shabandar et al., 2019, Gerlings and Constantiou, 2022, Bellomarini et al., 2020, Chen et al., 2018, Cirqueira et al., 2020]. The purpose of this explanation is to provide sufficient evidence about the suspiciousness of a case [Kute et al., 2021]. Gerlings and Constantiou [2022] investigated the needs for XAI in AML-CFT for banks' investigators and capacity planners. They highlighted the need to explain the

reasons for automatic closures of alerts and demonstrated the risk of bias when the scoring of an alert was made visible to the investigators.

However, very few studies have explored user needs from the perspective of supervisors. While Gerlings and Constantiou [2022] hypothesize that "auditors may require additional information on the model logic", they do not describe the supervisor's explainability requirements in more detail. Kuiper et al. [2021] explored the perspectives of banks and supervisors in the Netherlands regarding explainability in three financial domains, including AML-CFT. They found that supervisors expected explanations to have a broader scope than banking practitioners, who have a more technical and local understanding of explainability. They did not, however, detail the goals and needs of supervisors for explanations nor justifications and did not consider the legal requirements supervisors expect to see in model explanations.

6.3 *Methods*

This section presents the qualitative methods we used to understand the socio-techno-legal supervision system in AML-CFT and supervisors' needs for model justifiability. We first conducted five semi-structured, scenario-based workshops of two to three participants with 13 supervisors in total. At the beginning of our research, we had initially planned to study the need for transparency and explanation of the models, both for the supervisory authorities and for the banks, but we shifted our focus early on to the supervisory authorities in order to provide a more targeted and in-depth analysis. We nevertheless ran one workshop with participants from a large French bank, which improved our understanding of the existing supervisory mechanism from an other perspective: that of regulated entities.

During the workshops, we observed that the participants, particularly the supervisors, consistently referred to legal requirements or regulatory sanction cases when asked about the questions they had about the AI systems and the explanations or justifications they wished to see. This prompted us to find out more about the AML-CFT laws that participants referenced. Additionally, we noticed that the existing scientific or grey literature did not clearly indicate which legal requirements could undermine the use of AI. For that reason, we adjusted our initial research questions and added the RQ2 on how AI opacity conflicts with compliance requirements.

We present below the different methodological building blocks we used in the study, presented in chronological order of implementation. First, we present the procedure, artifacts used, and analysis for the workshops. We then present the methodology we used to complement the analysis of the workshops with regulation-driven needs for algorithmic justifiability. Lastly, we present our findings in post-analysis interviews with two experts in AML-CFT regulation.

6.3.1 Scenario-based semi-structured workshops

Procedure. All workshops were held in person at the participants' workplace and lasted between 90 and 100 minutes. Participants were not compensated. Upon their arrival, participants were asked to read and fill in a paper consent form. The consent form included a description of the purpose and possible risks (mainly confidentiality) of the study, the mitigating measures we implemented to ensure the confidentiality of the recordings and data presented in a publication, and finally their choice to voluntarily participate in this research and to be recorded. They were then asked to answer preliminary questions about their expertise in AML-CFT and their familiarity with AI on a printed form. The interviewer then detailed the workshop agenda.

The workshop questions focused on 4 main themes. First, participants were asked about the existing compliance procedure in AML-CFT in their profession (either controllers or bank practitioners). The following questions addressed the use of AI in AML-CFT to understand participants' impressions of AI. We originally planned this to find out more about how banking supervisors and practitioners envisage AI's future in AML-CFT. However, as the French supervisors were about to publish their position on AI at the time of the study, they considered this information to be too sensitive. We therefore limited the scope of our research to justifiability and explainability needs. We then presented participants with a scenario in which a supervisor controlled an AI-enhanced transaction monitoring system. We asked participants which kind of questions they had about the AI system and what kind of justifications they wanted to see. This scenario-based elicitation approach was used in prior research to understand users' needs for justifications and explanations [Liao et al., 2023, 2020, Sun et al., 2022, Rosson and Carroll, 2009, Wolf, 2019]. Finally, conceptual design artifacts [Gaver and Martin, 2000] of different explanations and justifications were presented to the participants for fictitious alerts. Participants were invited to discuss the relevance of the justifications and their limitations. As seen in Section 6.2.4, AI's entrance in AML-CFT is a recent topic where regulatory thinking has not yet matured. Therefore some of the questions called for speculative thinking. For this reason, we chose to interview the participants in small groups, so that they could discuss these issues together [Morgan, 1996].

Participants. One of the authors had several connections at the French Supervisory Authority to help contact the appropriate directors to obtain the necessary approvals to carry out the research and to connect with controllers. We also learned that the French Supervisory Authority has two departments, one for ongoing monitoring of all financial institutions registered in France and one dedicated to on-site inspections. We used the email lists for these two departments to recruit participants, describing the purpose of the research, the time, location, and agenda of the workshops. In total, we recruited 13 controllers from the French supervisory authority, 6 from the on-site inspections department and 7 from the on-going monitoring department. They had between 1 and 20 years of experience in AML-CFT supervision and their level of familiarity in

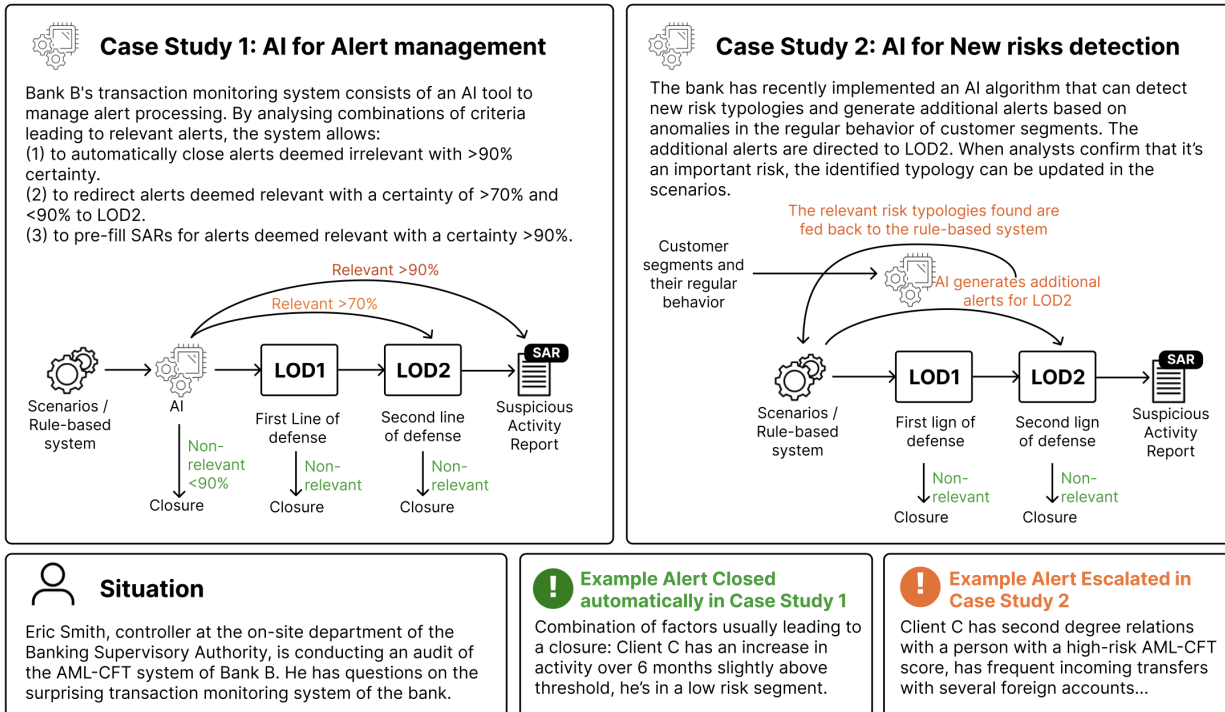
AI averaged 3.6 out of a Likert scale of 7; two participants had extensive expertise in AI—familiarity level with AI was 7/7.

The participants from the large French bank were recruited by a contact the authors had at the bank with a specific selection criteria for the participants, *i.e.* people specialising in AML-CFT with some previous exposure to AI and, if possible, also to supervisory compliance. In total, six participants took part in the workshop. Three participants’ expertise was AML-CFT compliance. The other three participants came from machine learning model development. Naturally, the participants in this study spoke in their individual capacity and their views do not represent the official positions of either the French Supervisory Authority or the Bank that employed them.

Of the 6 workshops, 4 were recorded and 2 were not as some participants did not feel comfortable with being recorded, notably due to the sensitivity of AML-CFT. However, participants who did not want to be recorded agreed to the interviewer writing notes. One of the unrecorded workshops was with controllers with extensive AI experience, the other was the workshop with banking actors. All participants were French and the quotes presented in this paper were translated from French to English by the authors. Table 6.1 details the profile of participants.

Figure 6.1: Scenarios used during the workshops with supervisors, with a description of the two use cases of AI in AML-CFT, and two examples of alerts that were generated or closed by the AI-enhanced systems. Only one of these case studies was presented in each workshop.

Artifacts provided.

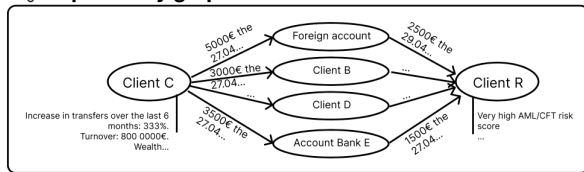


The scenarios featured a fictional character, Eric, whose role was either a controller carrying an on-site mission at a Bank B (for supervisors) or Bank B’s head of compliance (for banking practitioners).

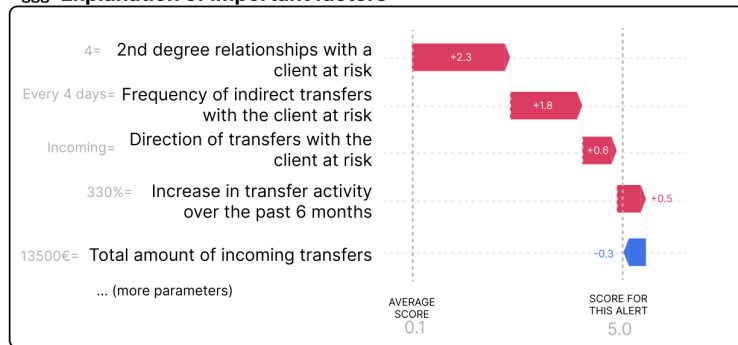
We designed two scenarios involving two types of AI-enhanced trans-

Case Study 2: Example of justifications for the escalated alert in LOD2

Explanatory graph of transfer activities



Explanation of important factors



? Probability of suspicion



AI system documentation

Training data	...
Role of the AI system within the existing system	...
Performance (#alerts reviewed per month, partial manual review of escalated/closed alerts...)	...
Performance gain compared to the system without AI	...
Choice of parameters	...
?	...

Explanation by example

"Here are the most similar cases in the bank's history. They all led to a SAR."

Certification

The following phases of the model building process have been certified by an external body:

- design
- development
- evaluation
- maintenance

Figure 6.2: Conceptual justifications shown for the scenario 2 and its example alert. Conceptual justifications for the scenario 1 followed the same format.

action monitoring systems which have been presented as the most common applications of AI in the scientific literature [Canhoto, 2020, Gerlings and Constantiou, 2022] and in reports from the French supervisory authority [Autorité de Contrôle Prudenciel et de Résolution, 2023b, Dupont et al., 2020]. In the first scenario, an unsupervised learning algorithm is used to detect new typologies of financial crime. This algorithm triggers alerts when it identifies a transaction as unusual for certain groups of customers that it has defined. Those alerts come in addition to the ones generated by the bank's traditional rule-based system, which generates alerts based on predefined rules or "scenarios", *e.g.* "transaction for this specific customer group is superior to \$10.000". When an alert is generated, a human analyst examines it and determines whether the identified risk should be addressed by the creation of a new rule in the traditional alert system. The second AI use case involved scoring alerts from Bank B's transaction monitoring system in order to prioritise, redirect, or close them. For high-scored alerts, a Suspicious Activity Report (SAR) was pre-filled automatically with generic information to be sent quickly to the Financial Investigation Unit. Only one scenario was used in each workshop. The first use case was used in three workshops and the second in the other three.

For each scenario, we described fictional example alerts triggered by the AI-enhanced AML-CFT system. For example, the example alert for the first scenario was an alert triggered by the unsupervised AI module. An example alert for the second scenario was an alert considered as low risk and closed by the AI. For these examples, we designed conceptual artifacts [Gaver and Martin, 2000] of different types of justifications and explanations. Our aim was to encourage participants to comment and imagine possible transparency solutions. We tried to balance the concreteness and openness of these artifacts and to leverage multiplicity in order to get feedback on the concept of these justifications rather than on their design. We chose to show the following justifications and explanations based on what we considered as most common in the literature on XAI for AML-CFT [Kuiper et al., 2021, Kute et al., 2021, Weber et al., 2023, Financial Stability Board, 2017].

- **a visualisation of the context** of the alert in the form of graph networks
- **a feature-based explanation** showing the most important variables for the AI-produced decision, their impact (positive or negative) and their weight
- **an uncertainty estimator** showing the probability of the alert to be suspect, as calculated by the algorithm
- **a model documentation structure**, including examples of sections: role of the AI system, training data used, performance evaluations, and choice of parameters.
- **an example-based explanation** presenting similar cases and their outcomes.

- **a certification** of the design, development, evaluation and maintenance of the model by an external body. We added this artifact because it is one of the provisions in the upcoming AI Act relating to high-risk AI systems.

Figure 6.1 presents the scenarios we showed to participants. The conceptual justification artifacts are presented in Figure 6.2.

Analysis. We used a content analysis methodology [Bengtsson, 2016] to analyse the audio transcriptions—including question-answering and think-aloud data—and the notes taken from the workshops. The notes were taken by the interviewer during the workshops and we recognise their limitations. Although they cannot reflect the details and nuances of the participants' thoughts and words, the notes nevertheless capture the general and sometimes strong opinions of the participants. The broad themes used for the content analysis followed the workshop structure: (1) the socio-technical context and (2) technical approaches of the supervisory authorities, (3) the AML-CFT legal requirements, (4) supervisors' questions on AI, (5) ideas for designing AI justifications and explanations. Based on the open codes gathered for each of these five overarching themes, we used axial coding to establish links between the concepts and refine them [Corbin and Strauss, 2014]. The first author, who was also the interviewer and note-taker for the non-recorded workshops, carried out the thematic and axial coding for 5 workshops—three fully transcribed and two partially-transcribed using notes. Another author analysed the audio transcripts of a workshop and applied open thematic coding separately. The two authors then discussed all the codes they had created and refined them on a Miro board⁶.

⁶ <https://miro.com/app/dashboard/>

6.3.2 *Empirical legal research*

As agents of regulation, supervisors' goals are embedded in the legal requirements they enforce. During the workshops, we observed that not having a full grasp of the various legal themes to which the participants were referring prevented us from capturing their motivations to ask for specific justifications. Therefore, we complemented the scenario-based eliciting approach with a qualitative empirical legal research [Webley, 2010]. We believe that combining needs elicitation with a legal analysis is key to fully understanding regulators' needs. In fact, the legal field is also keen on qualitative approaches, using interviews and legal document analyses, with methods similar to those used in the social sciences. Webley [2010] points out that "many common law practitioners are unaware that they undertake qualitative empirical legal research on a regular basis". We conducted this legal approach in parallel to the analysis of the workshops.

AI Compliance Assessment. Our methodology was adapted to address our research question, as recommended by Webley [2010]. It was carried out by the first author, who does not have a legal background, but the methodology and findings were discussed multiple times with another author with extensive experience in legal practice and research.

We began using a doctrinal research as described by McConville [2017], which consists in seeking what the law is in a particular area. We thus examined regulatory sanction cases on AML-CFT, the relevant articles of the French Monetary Code, and other useful legal documents on the advice of a lawyer from the French Banking Supervisory Authority. The data collected we used for this legal approach is detailed in Table 6.2. We narrowed our focus on AML-CFT and internal control requirements, as these are the requirements that banks are evaluated against during AML-CFT supervisory audits. We identified the main legal themes and specified their meaning, first using open coding on five regulatory sanction cases, because they reflect how supervisors' interpret and structure AML-CFT laws. We then refined the themes with the rest of the data collected. We used the scenarios we defined in Section 6.3.1 to assess how AI opacity impacts each identified theme. Finally we conducted feedback interviews. In short, our method follows these six steps:

1. Identify the applicable laws in AML-CFT and define the scope of the research through "*doctrinal research*"
2. Define the main themes in the applicable laws, building on the format of the legal documents and invoked themes in the workshops,
3. Specify the meaning of the requirements in each theme, drawing on the supervisors' perspective and legal documents such as case law, which inform on how the law is commonly interpreted,
4. Define scenarios featuring AI systems in AML-CFT,
5. Consider how the opacity of these systems conflicts with each sub-theme identified, which can also be formulated as goals for which the supervisors seek transparency,
6. Obtain feedback on our analysis from AML-CFT experts during interviews.

Feedback interviews. Because step 5 of the above methodology can be somewhat subjective and potentially inaccurate due to the lack of expertise of the first author in AML-CFT law, we conducted two interviews to elicit feedback and corrections from experts. The two participants were solicited upon advice from internal contacts at the French supervisory authority, given their unique expertise in both AI and law. One of them was a lawyer and the other an on-site inspector with extensive background in AI. Our pre-interview included a presentation of the research, confidentiality risk mitigation measures, and request to record interviews. We began by asking participants two general questions: what do they see as the key challenges in assessing AI's compliance with AML-CFT requirements, and how does the opacity of AI make compliance with AML-CFT requirements difficult. We then presented them an initial version of the table shown in Appendix C2 and asked for feedback. Interviews were used to both correct and complement our prior analyses. Interviews were recorded, transcribed, and two authors analyzed and coded them according to the process described in Section 6.3.1.

6.4 Results

The results presented in this section are structured around three axes, each aimed at improving our understanding of a user group that is under-represented in the literature: regulators, more specifically, supervisors in AML-CFT. The three axes correspond to our research questions: understanding the supervisors' socio-technical context (RQ1), understanding the regulatory goals of supervisors in AML-CFT (RQ2), and articulating the supervisors' needs for AI justifications and explanations (RQ3).

6.4.1 Socio-techno-legal context and auditing approaches of supervisors in AML-CFT

Figure 6.3 provides an overview of the workshop findings and the socio-techno-legal context of supervisors.

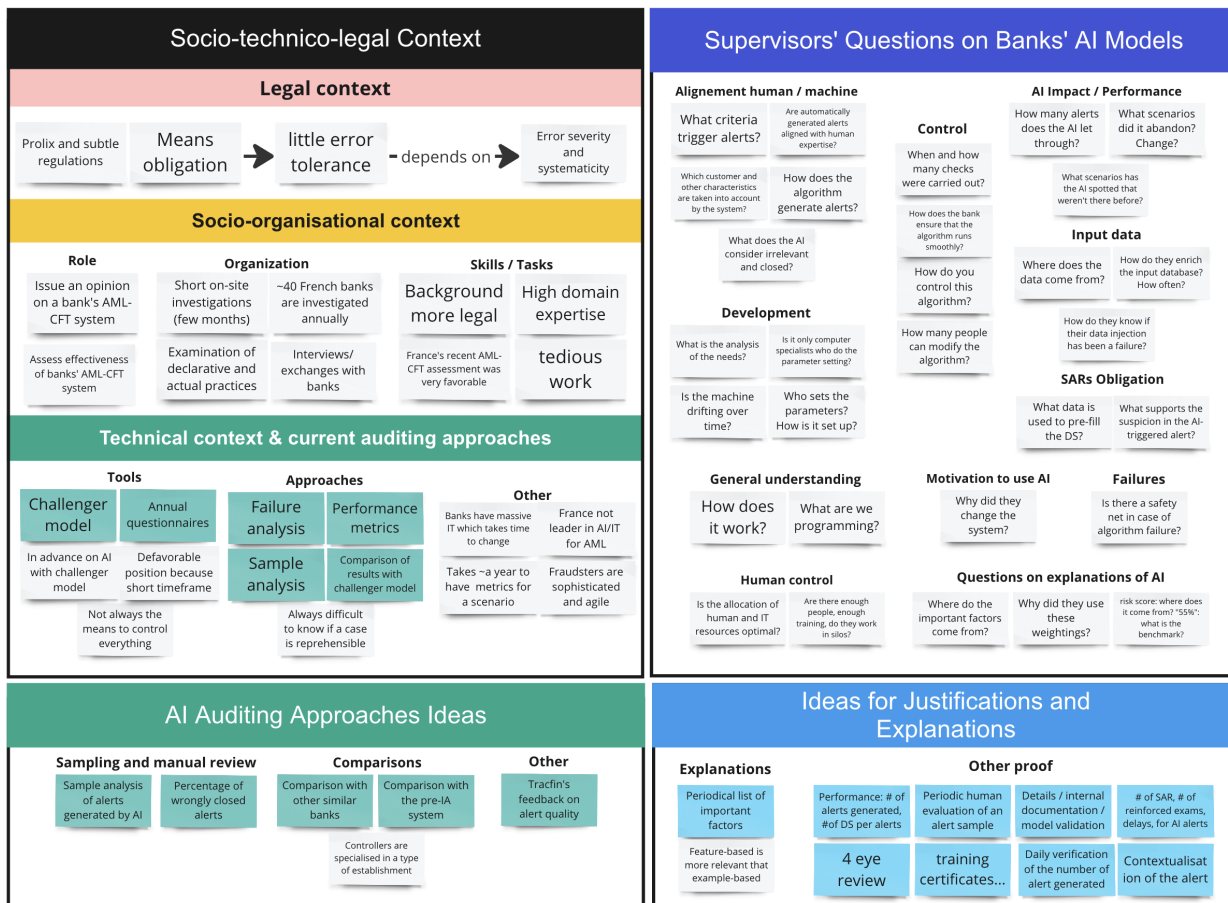


Figure 6.3: Summary of the workshops, with socio-techno-legal context of supervisors, supervisors' questions on AI, AI auditing approaches ideas and ideas for justifications and explanations.

How are supervisory audits organized in practice?

The French Banking Supervisory Authority carries out two types of

inspections: document-based control and on-site.

The document-based control unit's mission is to **assess the maturity of the AML-CFT system of each regulated entity in France (around 1,300)**. This control is based on numerous records, including an AML questionnaire that banks report annually and exchange with the regulated entities. They then notify the banks of their observations. This unit can also suggest on-site inspections, as one participant notes:

"when we see a lot of deficiencies, we will inform the on-site inspection and propose that the establishment be included in the investigation programme".

The role of on-site inspections is to **confirm the true state of a bank's declarations concerning their system for AML-CFT and to assess their effectiveness**. Inspectors will challenge a bank's system, observe how employees work, compare declarative practices with what actually occurs, exchange information with bank practitioners, and perform IT extractions to identify any major deficiencies within the allotted time for inspection, *i.e.* a few months. One participant emphasised the importance of the iterative process when communicating with banks which helps prevent misunderstandings. Around 40 on-site investigations take place annually [Autorité de Contrôle Prudentiel et de Résolution, 2023a]. Following the findings of an on-site inspection, a sanctions committee may then be called upon to decide whether a penalty should be imposed. Figure 6.4 details the anti-money laundering and terrorist financing controls for the French supervisor.

It is worth noting that the large majority of controllers have a legal background with expertise in financial crime analysis. Many participants, therefore, expressed unease with complex statistical tools such as AI. For example, some participants said *"our IT skills are a little limited"* (P3) and expressed their lack of computer science knowledge to deal with the particularities of machine learning models. One of these participants, however, was aware of unsupervised and supervised learning and many participants with little familiarity with AI were able to generally describe the functioning of the AI-based systems they had seen in banks. Moreover, on-site missions include at least one computer scientist to support non-tech controllers. One participant stated

"When you need to go into details, you need to have knowledge, experience or even ideas of what to do. Their [the banking actors'] job and ours is evolving, we'll have to speak both the financial crime and python languages." (P11)

How do supervisors describe the legal context in AML-CFT?

Section 6.2.4 provided an objective review of the legal context. Below we give a brief impression of participants' perspectives on these regulations. Supervisors described the AML-CFT regulation as *"prolix"* (P1) and *"subtle, with high expectations and not much room for error"* (P11). Another participant added that *"every system, even the best, does not detect everything"*, confirming that a **small margin for errors** is left in transaction monitoring given there is an obligation of implementing the best means and not an obligation of results. Just as there exists a small margin for error for

data quality⁷ they expect AI tools to also make errors. Supervisor tolerance is qualitative, and depends on error severity and systematicity. It was also noted the regulation does not stipulate a requirement to automate tools. It is instead the size of the regulated entity and its volume of transactions that will drive an implementation of automated "scenarios" and ultimately, AI. One participant noted that

⁷ roughly below 5%

"[Banks] are fairly up to speed with regulation, they will end up on AI one day or another."

What are the approaches of supervisors to audit the automated AML-CFT systems in banks?

Participants emphasized that there is no single approach to auditing; all audits adapt to their context. We identified, however, some common approaches to auditing. Investigations or document-based assessments usually start by examining the risk classification of banks⁸. Banks must produce this document, which identifies the money laundering and terrorist financing risks related to the bank's activities, size, customers, etc. Supervisors can then identify gaps in the identified risks, in the risks covered by scenarios, and other automated tools. Then, during controls, supervisors assess the quality and compliance of two aspects of the bank's AML-CFT systems: processes and results. Approaches to evaluate results may pinpoint failures in the process and vice versa. Audit strategies of AML-CFT frameworks can be broadly summarized in three approaches: "global", "global to local" and "local towards global".

⁸ One participant noted: *"everything flows from the risk classification"*

Global approaches consist in looking at metrics characterising the efficiency of AML-CFT devices. These metrics include, for example, the number of alerts generated, the number of reinforced examinations, and the number of SARs. Supervisors interpret these metrics in relation to the bank's characteristics; as a participant notes,

"We'll see if they're consistent with the establishment's activity." (P3)

It takes some time, however, for these measures to reflect the value of a new tool:

"as long as the scenario hasn't really run for a year, we won't have very interesting statistics." (P4)

Furthermore, a **"global to local approach"** enables controllers to find cases to investigate. The French supervisory authority recently developed an AI-based tool, "LUCIA", to support controllers in sampling cases and comparing them with the bank's results [Laporte, 2021]. Participants highlighted time-savings and novel offerings of this tool:

"It makes it possible to review, I don't know, thousands of operations, whereas as an on-site controller we can see a panel of about fifty operations." (P8)

P1 reported that the work of controllers is often very tedious and stressed the need for tools like LUCIA,

"so that we are in a position, not to anticipate anything, but to react to regulations and perhaps to detect loopholes more easily." (P1)

P7 summarized the main goal of SupTech tools:

"enrich the control by giving possibilities or ideas that the analysts would not have had or that they would not have had the means to look at." (P7)

Local approaches involve examining specific cases or part of the AML-CFT framework to see if there are any crude errors in reasoning. Examining local cases can also give conclusions about the results. **The "local towards global" approach aims at drawing conclusions on the system from ad-hoc observations.** Supervisors draw on a thread of errors observed in specific cases to trace systematic errors in the system. This is enabled by "failure analyses" or "sample analyses" which consist of examining cases either brought to the attention of supervisors by TRACFIN or another public authority, or drawn from a sampling strategy. Supervisors ask:

"should the system have detected [the errors]? Was it within its scope? Was it within its objectives and why didn't it detect them, what went wrong? " (P14)

Overall, **the superposition of different methods** for auditing and detecting financial crime in banks, whether AI-based or not, improves the efficiency and robustness of the frameworks:

"We know that there will be illegal operations that go undetected. We can't detect everything, but there's an obligation to try and detect as much as possible, and if we start relying solely on AI, well, we're bound to miss things. But we'll miss less if we superimpose different methods." (P14)

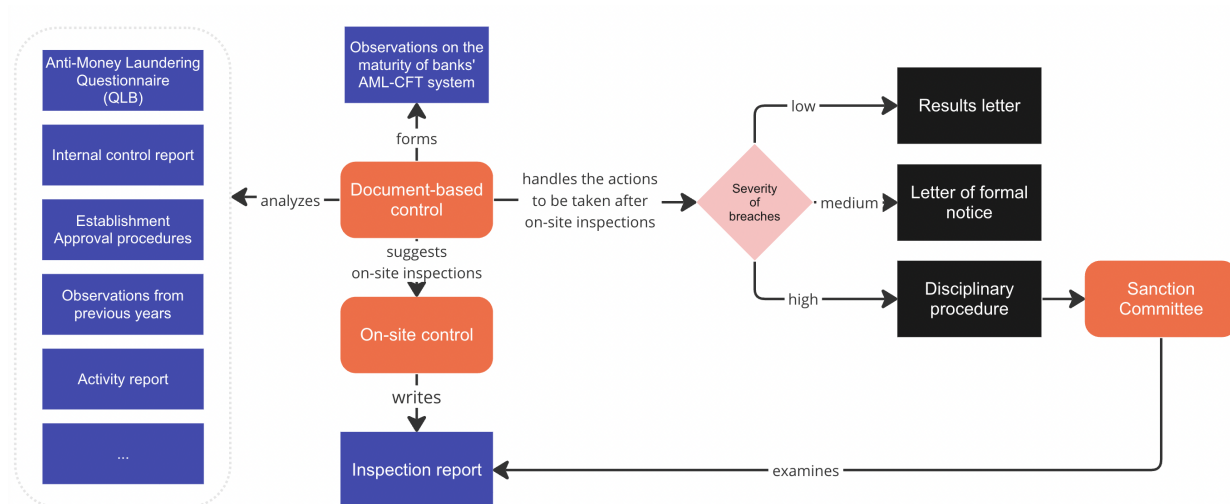


Figure 6.4: Flow diagram of the supervisor's control procedures in AML-CFT

6.4.2 *What provisions in AML-CFT laws does AI opacity conflict with?*

This section presents the results of our compliance assessment, the methodology of which was presented in Section 6.3.2. The paragraphs below present a regulatory goal (RG) with which AI opacity can conflict. Table C.2 in the Appendix also provides a summary of this analysis.

Verifying risk adaptation (RG1)

As part of compliance requirements, supervisory authorities verify the adequacy and completeness of a bank's operation monitoring system in relation to its risk classification⁹. Much of this assessment is based on a qualitative understanding of the reasoning and criteria used by the system to generate alerts. This enables controllers to verify that important characteristics of the business relationship are considered (e.g., income), or that the thresholds are relevant based on business expertise. The opacity and complexity of AI led some participants to fear that this assessment would become difficult:

"We're going to end up with this like chickens with a knife and we won't know exactly why it generated this alert...we won't be able to assess the adaptation to the risk". (P4)

⁹ c.f. Article R. 561-12-1 of the French Monetary Code (CMF) and Decision against AXA Banque of the 15/02/23

Verifying the bank's ability to perform constant and careful examination (RG2)

Supervisors also have to verify that transaction monitoring systems detect inconsistencies with up-to-date customer knowledge and fulfill the bank's obligations of carrying out "careful examinations" of operations¹⁰. Supervisors typically use performance metrics and a "local to global approach" to evaluate this. As AI algorithms are opaque, however, supervisors may not be able to establish if an ad-hoc error in detecting financial crime is linked to a broader issue in the system. Moreover, clarifying how AI systems adjust to input updates might be needed to comply to constant vigilance obligations.

¹⁰ c.f. Article L561-6 of the CMF

Verifying the bank's ability to perform "enhanced vigilance", to produce quality Suspicious Activity Reports, and to update their risk classification (RG3)

Financial institutions also have the obligation to increase surveillance with regard to complex or risky transactions and to submit high-quality SARs to TRACFIN. As one participant said:

"All alerts must be duly substantiated and analysed." (P10)

This implies that sufficient explanations be given on why a scoring algorithm (as in the first scenario) considers an operation as risky and why an alert was generated by an algorithm (as in the second scenario), so that human analysts can write high-quality SARs:

"We need to be able to understand the criteria that generate a risk. It's a question of auditability. Actually, before that, it's a question of a human analyst's ability to understand what to look at." (P14)

Verifying that banks can detect incidents and have control over the purpose and operation of any device used (RG4)

Internal control obligations require banks to: be able to detect incidents; control the operation of their devices, notably over time; demonstrate control over the purpose of their system, particularly when it is provided by a third party; and plan for safety nets in case of failures¹¹. However, AI opacity can prevent banks from correctly **detecting instabilities like drift or anticipating failures**:

"If you don't know what behaviour is expected, you can't say that there's been a malfunction." (P10)

The inscrutability of algorithms can also create **dependencies on AI**:

"there is a risk of dependence on AI if the criteria are not understood." (P7)

Verifying the correct allocation of material and human resources (RG5)

AML-CFT laws also require banks to put in place the material tools and human resources needed to monitor operations¹². Case law indicates that it is a question of striking a balance between human and automated tools. AI transparency will be needed to **show how human expertise and AI systems are balanced and complementary**. Many participants insisted that human expertise cannot be replaced in many instances:

"there is a human expertise that cannot be replaced, particularly in advising banks on signs of radicalisation..." (P1)

For that reason, the auto-filling of SARs by AI, if not verified and substantiated by a human, as presented in scenario 1, was seen as problematic. Moreover, explainability can have a major role in enabling transitions between machine and human analysts and to ensure timely processing of the alerts, as P10 noted:

"there may be an impact of explainability on processing times."

Indeed, SARs should be filed without delay so that TRACFIN can bring cases to court as quickly as possible.

Understanding the motivations for AI use (RG6)

Some participants, during the semi-structured workshops, were also questioned on whether banks needed to justify the use of AI. Most participants claimed that while it is not legally required, it could help better understand the implemented transaction monitoring system. One participant explained:

"I'd use motivate rather than justify, in other words, the Bank is free to use AI. On the other hand, it must always be able to motivate, to explain why such change in its system." (P7)

¹¹ C.f. Article R561-38-4 of the CMF, Order of November 3, 2014

¹² c.f. Article R561-38 CMF

6.4.3 Supervisors' needs for model justifiability in AML-CFT

The summary of the workshops presented in Figure 6.3 shows the questions that supervisors asked about the AI systems described in the scenarios. Based on the supervisors' regulatory objectives described above and their questions about AI, we formulate supervisor needs for justifiability below.

Understand the basics (N1)

Supervisors who are primarily lawyers require high-level explanations or machine-learning training to answer their questions like "How does it work?", "What are we programming exactly [in machine learning programs]". They want to be able to autonomously use a "Challenger" model, *the supervisor's AI model*, to assess bank's systems. As noted by one participant:

"controllers have to be able to understand the purpose and operation of the SupTech tools that their IT team implements" (P11)

Their profession will evolve towards hybrid profiles that are both legal and technical. However, the current challenger model developed by the Supervisor, LUCIA, is designed as a support tool for in-depth analyses. One participant explained:

"Paradoxically, the stakes may not be so high because you get to the stage where you're digging into the details anyway, and then you abstract from the surveillance system." (P10)

Demonstrate legitimacy (N2)

With LUCIA, supervisors are in an advanced position where AI is challenging traditional rule-based systems. The errors found during this process also highlight the added-value of AI, one participant noted. However, participants from the bank have stressed the **need to be on a level playing field, according to the legitimacy principle of due process rights of regulated companies** ("equality of arms") [OECD, 2021a]. For that purpose, they would like to understand the data or methodology used by the supervisor, especially data they do not have access to. Banking professionals also wanted to know if the challenger model was using sensitive data, or if it was discriminatory in any way, as they are entities subject to privacy regulations¹³. Nevertheless, a supervisor pointed out that they are rather at a disadvantage when it comes to finding undetected financial crime, which fuels their need for AI tools:

"the tight time-frame [for investigations four months]¹⁴, we need to start everything from scratch each time, the data, everything..." (P14)

Supervisors have implemented question-answering sessions for banks on this issue.

¹³ The participants from the bank were concerned that LUCIA would use insights coming from comparisons with other banks or sensitive data, but this is not the case. The AI-based supervisory tool only relies on the data provided by the inspected bank [Laporte, 2021].

¹⁴ which is already longer than in some other countries, where they investigations are sometimes carried out in a flash (a few days), the participant noted.

Measure global efficiency (N3)

The global approaches described in Section 6.4.1 to measure the AML-CFT framework performance are likely to remain valid for any system, AI or not. One participant indicated that "*Even before AI, the black box phenomenon already existed.*" (P14). In particular, the current sampling strategy by the supervisory authorities is still suited to assess AI-enhanced AML-CFT systems.

"For us, the most practical and realistic way of checking that this [the system] is not absurd is not to look at the parameterisation. Because it's difficult to understand the effects of a parameter when it interacts with other parameters. It's a question of seeing in situ how it behaves in reality when faced with examples that we have selected ourselves." (P14)

A participant indicated three main approaches envisaged for evaluating global performance of AI-enhanced AML-CFT systems: (1) compare efficiency with the pre-AI system, potentially comparing performances with similar establishments; (2) analysis of the "failures" reported to the supervisory authority; (3) comparison of the banks' results with the results obtained using a challenger model on sampled cases. The sampling approach was mentioned in all the workshops with supervisors.

P1 and P2 also brainstormed about "*simple, basic*" indicators to measure efficiency, using, for example, the ratio of suspicious transaction reports to turnover "*or something similar*", refined for relevant clusters of similar establishments, potentially made with AI. Aggregated statistics of **this indicator could also be shared with financial institutions to encourage improvement:**

"if we give them the average, they set themselves a performance target which is, I don't know, like, 20% above average." (P2)

Another group of participants felt more dismayed by the increasing opacity and complexity of AI systems. They argued for another approach to measure efficiency that relies more on financial intelligence units:

"the standard controller will be completely helpless. We'll have to change the way we monitor, we'll have to work more with the financial intelligence unit, TRACFIN, which will then be the only one able to give an opinion on the alerts."

Establish reprehensibility (N4)

Despite implementing sampling strategies, having a closer look into the AI system inner workings might be necessary to establish the reprehensibility of the errors detected. Understanding why a suspicious transaction was not detected might help conclude on the systematicity, and therefore the reprehensibility of the problem. This requires a contrastive explanation, focusing on the negative which answers questions such as "**why did the system behave in this way (letting the fishy transaction go) and not in this other way (flagging the transaction)?**". One participant described:

"It's the question of how you go from analysing individual declarative failings to making structural observations about the structural failings of the system." (P10)

Banks also need to implement such explanations when implementing anomaly detection AI systems, as in Scenario 2. In this case, the unsupervised algorithm may encounter a risk typology, not covered by the traditional bank's system. The bank then has to understand why this risk was not detected and, if necessary, update the risk classification.

Verify and challenge banks' AI understanding (N5, N6, N7)

As noted in Section 6.4.2, supervisors may need to examine a bank's explanatory practices to **ensure that analysts are able to understand alerts** and justify their suspicious nature (N6). To that end, justifications based on local feature importance explanations, which would be implemented by banks, have been preferred by participants:

"the feature importance explanation is more interesting than the example-based one, which is quite limited eventually." (P7)

Bank participants said they were currently testing an explanation based on Shapley values [Lundberg and Lee, 2017]. The contextualisation with graphs networks has also been appreciated by some participants. In the advent where graph neural networks would be used, we can also imagine that graph visualisation will be highly recommended by supervisors, as is the case for digital asset service providers using blockchain, one participant commented. Views regarding uncertainty estimators were divided. One participant mentioned that:

"It is important to know whether the connections made are coincidental or not." (P14)

. However, some participants warned against the confirmation bias it can trigger:

"all these very precise indicators create a push-button risk: as soon as there's a lot of red, bang! [the alert is escalated]." (P9)

Bank participants also confirmed they saw investigators fall into this bias when testing explanations.

Supervisors also want to **verify the human alignment of the decision criteria** used by AI systems (N6). Even though the need for explanations of supervisors is more global, they may look for ad-hoc examples of local explanations:

"We're more interested in the global [...] We'll ask them for local, but local examples for specific cases." (P7)

Supervisors will not only be interested in the explanation, but more importantly in the justification of why or how developers have validated these feature weights:

"The weight has to be less than..., OK a priori, but why?" (P6)

"It can be a relatively aggregated explanation, i.e. we're not trying to go into the details of the calculation, but to identify the main steps." (P8)

Finally, supervisors also need justifications that **banks control what their AI system is doing** (N7):

"it's the idea that it creates a dependency on the AI and that the day the AI changes or is hacked, we don't notice the change because we don't know what was at the origin?"

Feature-based importance was seen as useful to that goal:

"with the feature importance explanation, we'll be able to assess: are we in agreement with all these factors?" (P7)

Another participant mentioned that justifications, such as the **daily number of alerts generated, and periodic human verification of a sample of alerts** could be effective measures to prevent drift. **Documentation was also seen as crucial** for N7 and N6: *documentation is super-important to check that they master their tools (P9)*. Certifications from third parties, however, elicited more cautious responses. Some supervisor participants argued that, if certification was to become the norm for AI models, it would put regulators in the difficult position of having to adjust the scope of their audits. Other participants from the AML department of the supervisory authority said they would ignore this third party accreditation which infringes upon their role.

6.5 Discussion

In this section, we discuss the importance of relying on accurate information about AI systems to justify compliance, explanations' limits and alternative approaches like tests or challenger models.

6.5.1 The role of explanations for justifications

In this paper, we saw that regulators mainly seek *justifications* from regulatees, *i.e.* argumentative demonstrations that their AI systems comply with certain legal requirements. Justification is therefore a critical element in the process of enforcing regulations, *i.e.* for auditability and more broadly for accountability [High-Level Expert Group on AI (HLEG), 2019]. Just like explanation, justification is a process [Miller, 2019]. One participant mentioned the importance of exchanging with regulatees. Another mentioned that *"justifications are meant to be challenged"* (P11).

[Henin and Le Métayer, 2022, Hamon et al., 2022, Hildebrandt, 2019] argued that explanations are not sufficient to justify a decision. Further, Hildebrandt [2019] added "we must not allow the discourse of explainability to stand in the way of the question whether a decision is legally justified, which requires a specific type of legal reasons" [Hildebrandt, 2019, Henin and Le Métayer, 2022]. Additionally, Henin and Le Métayer [2022] precise that "justifications are complete only if they establish a continuous link between the high-level objectives of the [AI system] (the applicable norms, for example non-discrimination, reduction of recidivism rate, or compliance with a given legal requirement) and its implementation". The authors also stress that justifications are "extrinsic" in the sense that they refer to external norms such as legal requirements.

However, we argue that acceptable justifications about AI systems should also take into account descriptive, intrinsic, and accurate information about the "implementation" of AI models, to establish this "continuous link". Just like explanations may not always be sufficient to ensure the legitimacy of AI systems, information about an AI system's objectives, design choices, or performance may not always be sufficient to justify the proper implementation of AI models. Furthermore, justifications are intended to be challenged and if they do not rely on factual information about algorithms, there is a risk that the question of the legitimacy of an AI system becomes subjective and arbitrary. In their paper about algorithmic audits, Koshiyama et al. [2021] argued that, without explainability, a decision cannot be duly contested. Explanations may therefore be insufficient, but are necessary, to provide descriptive, accurate and faithful information about the behavior of an algorithm on which to develop a justification.

The list of needs described in Section 6.4 illustrate why regulators may need justifications from banks in AML-CFT, whether those rely on explainability or on other kinds of proof such as documentation or tests. In AML-CFT, regulators not only assess results but also processes. Therefore, looking at explanations of the inner workings of AI systems, even high-level ones [Bibal et al., 2021, Dupont et al., 2020], may become necessary, not only for banks but also for supervisors. The needs N₁, N₂ and N₄ in Section 6.4 reflect this.

6.5.2 *Considering the limits of explanations*

However, current XAI techniques may fall short of regulators' expectations to provide accurate and faithful information about AI system's inner workings. As outlined in [Hamon et al., 2022, 2020], the fidelity, robustness and truthfulness of explainability can be limited by the fact that the many features used by complex algorithms are highly correlated. This is a well-studied and strong limitation of feature-based explanations, which make it difficult to comply with legal requirements to indicate the most important factors in a decision [Hamon et al., 2022, Rouvroy, 2013]. This goes back to the question of the reliance of AI systems on correlations rather than causal relationships. This can be an issue for measuring model performance as well [Hamon et al., 2022].

Another issue with explanations is that they can be misinterpreted by their users due to the technical language they usually use. Ronan et al. call it the "transparency fallacy" when explanations are not effectively understood. We saw this in the reaction of some of the participants in this study, who were unsettled by the precise weightings given by the feature importance explanations. Moreover, as demonstrated by Gerlings and Constantiou [2022] and highlighted by some participants, investigators must have access to sufficient information other than explanations, specifically risk scores, or they will fall into confirmation bias. Supervisors will therefore need to verify that the context in which explanations are presented to investigators, or supervisors themselves, takes account of this bias and mitigates it.

Given their mostly legal background, regulators may also be too quick to accept these explanations as trustworthy. Moreover, the argumentative process of transforming explanations into justifications could be used to the advantage of regulated entities to conceal technical inaccuracies. For example, Zhou and Joachims [2023] investigate the concept of "malicious justification". They develop a malicious explanation system that replaces the discriminatory factors (*i.e.* race) used by a biased decision model with other, non-discriminatory factors to defend the decision. Further, they demonstrate that it's almost impossible even for auditors, who have access to all the decisions, to uncover the deception. The authors also highlight that current explanations do not provide answers to questions like: "what factors caused the model to predict X instead of Y?". Yet, as highlighted in Section 6.4.3, supervisors are likely to need such contrastive explanations to establish reprehensibility of failure cases (N₄). As a result, regulators may be in a difficult position to evaluate the adequacy of explainable methods developed by banks, and may have to develop their own "explainability challenger" toolkit.

Lastly, Lima et al. [2022] argues that there is a trade-off between accountability and explainability, stating that post-hoc explanations such as feature-importance could "obscure the responsibility of developers in the decision-making process". While this phenomenon might be mitigated in highly-regulated industries where solid accountability mechanisms are in place, it is worth bringing this to the attention of regulators.

6.5.3 *Supporting model performance measurement and testing*

To address the limits of explainability to audit AI systems, specifically regarding fairness, Zhou and Joachims [2023] suggest that system-wide metrics are more useful. This was overall supported by the supervisors interviewed in this study. In fact, system-wide evaluation is a pillar in the auditing approaches implemented by the AML-CFT supervisor. This is reflected in the role of the document-based unit: assessing the maturity of banks' AML-CFT systems, and in the new challenger model developed for investigations. Supervisors are therefore more likely to continue on that "global" or "local to global" path, *c.f.* Section 6.4.1.

In the field of AML-CFT, however, current metrics to evaluate the effectiveness of systems are limited, notably because banks and supervisors, do not know the ground truth regarding alerts, *i.e.* whether a suspicious case was actually money laundering or not. Instead, they have to rely on proxies such as number of suspicious activity reports. The supervisor may have more feedback on the ground truth through the financial investigation unit, but perhaps not to the point that they can calculate the precision of the system, *i.e.* true positives reported to the sum of true positives and false positives. AI's entry in the industry could represent an opportunity for the supervisor to get closer to the financial investigation unit, as one participant noted.

The consolidation and disclosure of aggregated data such as precision on the performance of AI models from different banks could be useful for

the regulated entities self-assessment and research purposes. In health-care, the disclosure of a database of AI-based medical technologies with regulatory approvals enabled researchers to point out some AI weaknesses [Meskó and Topol, 2023]. Further, such initiatives can help respect the due process rights of regulated entities (N2), while striking a balance with advancing the fight against financial crime.

However, this approach does not inform on the false negatives of AML-CFT systems. Challenger models such as LUCIA can do this to some extent by identifying some crimes that have fallen through the cracks. However, they cannot fully measure the true proportion of crime that has not been detected. This calls for relative comparisons instead of absolute ones, such as comparing banks' practices or pre-AI systems as outlined by participants.

Lastly, to verify processes in addition to results, supervisors in this study have proposed some testing and human oversight mechanisms. More advanced testing methods will however have to be developed to prevent risks specific to AI such as drift, discrimination, over-reliance on AI. Certifications of the model development were seen as overlapping with supervisors' role. Discussions between certification providers and supervisors might be beneficial to talk about best practices, such as standard models for documentation [Mitchell et al., 2019, Gebru et al., 2021], or mathematical proofs that a code is correct, when applicable [Henin and Le Métayer, 2022].

In summary, future work could investigate the design of:

- contrastive explanations to help supervisors establish reprehensibility of failure cases (N4),
- meaningful sectorial, system-wide, metrics and databases to compare the efficiency of AI-enhanced systems in relation to each other or to pre-AI systems (N3),
- meaningful tests for AI to support supervisors in verifying correct use of XAI (N5), human alignment of decision criteria (N6) and model drift control (N7).

6.6 *Limitations*

As the scenario-based elicitation task came fairly early in supervisors' thinking about the use and audit of AI, their responses may not include in-depth considerations on the issue. The purpose of this paper was to articulate the needs of supervisors at a time when the use of AI in AML-CFT and investigations into AI-enhanced systems are in their infancy. We recognise that their needs may evolve as AI audits in AML-CFT develop and new regulatory and case law guidance is issued. Moreover, our research results rely on the specific scenarios and artefacts we presented to participants. This may limit the scope and generalisability of the results. Specifically, we investigated two use cases of AI, which are considered as the most common and promising in the literature, but other AI applications exist [Chen et al., 2018]. We also limited the number of conceptual explanations and justifications to six to not overwhelm the participants and to respect their time as volunteers. Other explanations could be considered in future explorations with regulators. Further, we described in the methodology section that two workshops were not recorded due to participants' concerns, we are aware that this limits the analysis and findings from those workshops. However, we were able to conduct a recorded interview with one of the participants in an unrecorded workshop, which enabled us to study the views of this person more closely. Finally, as the first author who conducted the legal approach has no legal training, the method remains fairly straightforward, but we did put in place quality controls with another author, who has a legal background, and two AML-CFT experts. We hope this study demonstrates the feasibility and suitability of such an approach for HCI practitioners.

6.7 *Conclusion*

In this chapter, we examined a socio-techno-legal supervision system in a highly-regulated industry, taking the example of the anti-money laundering and countering terrorism financing domain (AML-CFT) in France. We drew on 6 workshops with supervisors and bank practitioners to outline the auditing approaches of AML-CFT supervisors. We then outlined AML-CFT compliance requirements which raise clear issues with AI opacity, and drew up a list of seven model justifiability needs for the supervisors, integrating explainability aspects. In particular, we found that supervisors primarily need to measure the performance of the AI-enhanced AML-CFT system. However, supervisors may need contrastive AI explanations to establish the reprehensibility of sampled failure cases, to verify and challenge banks' correct understanding of the AI and to demonstrate the legitimacy of their challenger model. These needs are intricately linked to the regulations that supervisors enforce, hence the need for a dual interview-based and legal approach. We also presented explanations as having a role of "trial evidence" for justifications. We hope that this work will inform future research to design AI justifications for regulators.

Participant ID	Role		Years in profession	Familiarity with AI (on a 7 points Likert scale)	Workshop and Interview ID	Recorded
P1	Supervisor, document-based control		>10	2	W1	Yes
P2	Supervisor, on-site control		>10	3	W1	Yes
P3	Supervisor, document-based control		Between 1 and 3	3	W2	Yes
P4	Supervisor, document-based control		Between 4 and 10	3	W2	Yes
P5	Supervisor, document-based control		Between 1 and 3	3	W2	Yes
P6	Supervisor, document-based control		Less than a year	3	W3	Yes
P7	Supervisor, document-based control		Between 4 and 10	5	W3	Yes
P8	Supervisor, document-based control		Between 4 and 10	3	W3	Yes
P9	Supervisor, on-site control		Between 1 and 3	7	W4	No
P10	Supervisor, on-site control		Between 4 and 10	7	W4, I1	No, Yes
P11	Supervisor, on-site control		Between 4 and 10	1	W5	Yes
P12	Supervisor, on-site control		Between 4 and 10	3	W5	Yes
P13	Supervisor, on-site control		Between 4 and 10	3	W5	Yes
P14	Supervisor, AML-CFT policy		>10	6	I2	Yes
P15	Bank, Head of AML-CFT compliance		>10	3	W6	No
P16	Bank, Head of data science		Between 4 and 10	7	W6	No
P17	Bank, AML-CFT Compliance Officer		Between 4 and 10	1	W6	No
P18	Bank, AML-CFT Compliance Officer		Between 4 and 10	3	W6	No
P19	Bank, Data scientist		Between 1 and 3	7	W6	No
P20	Bank, Data scientist		Between 1 and 3	7	W6	No

Table 6.1: Description of role, experience, familiarity with AI of participants in the study.

Type	Document
Regulatory sanction cases	<ul style="list-style-type: none"> • Sanction Commission Decision 2022-04 against BMW Finance • Sanction Commission Decision 2022-02 against Financière des paiements électroniques • Sanction Commission Decision 2022-01 against Axa Banque • Sanction Commission Decision 2021-05 of 1 December 2022 against Caisse régionale de Crédit agricole mutuel du Languedoc • Sanction Commission Decision 2021-01 of 1 March 2022 against W-HA
Law, orders	<ul style="list-style-type: none"> • AML-CFT: Articles L561-1 to L564-2 of the French Monetary and Financial Code [Légifrance, 2023b] • Internal control: French Monetary and Financial Code, Articles L511-55, L522-6, L522-14 and L526-27, Order of November 3rd, 2014 [Légifrance, 2023a].
Soft law	<ul style="list-style-type: none"> • Joint ACPR and Tracfin guidelines on reporting obligations to TRACFIN • Thematic review: Automated systems for monitoring of AML-CFT transactions
Interviews	<ul style="list-style-type: none"> • 5 Workshops with 13 supervisors/controllers • 2 Interviews with 2 AI/AML-CFT supervisors

Table 6.2: Data used for the empirical legal research

Need	Description and related regulatory goal	Model / XAI Developer	Design ideas for explanations and justifications
N1: General comprehension	Understand how the challenger model works to extract relevant and representative case samples. Have a general understanding of how the banks' algorithm works (RG6).	Challenger and Bank model / Supervisor and Banks	High-level and global explanation, practice using the model and training, descriptions and motivations of AI's role
N2: Ensure legitimacy and efficiency of challenger model	Monitor performance of the challenger model and make banks appreciate the overall workings of the challenger model.	Challenger model / Supervisor	Global explanation, specific question-answering with banks
N3: Measure efficiency	Measure the performance of the algorithm, not only in absolute terms but also more concretely in a relative way. Linked to (RG1), (RG2), (RG3).	Bank's model / Bank and Supervisor	Performance metrics: delays, number of SARs, number of reinforced examinations, sampling analysis, Tracfin's feedback on alert quality
N4: Establish the reprehensibility of sampled error cases	Understand why a bank's algorithm did not detect a suspicious case, so as to understand if it was an isolated event or part of a bigger pattern: is the error systematic, reprehensible? Linked to (RG1), (RG2), (RG3).	Bank's model / Supervisor	Local feature importance, Counterfactual explanations
N5: Verify correct use of explainability	Ensure that banking analysts have a clear understanding of the alerts they are required to handle, so that they can produce high-quality analyses. Linked (RQ3), (RQ4), (RG5).	Bank's model / Bank	Justifications that explanations for analysts are present and efficient, alert contextualisations
N6: Verify human alignment of decision criteria	Verify that the criteria used by AI to generate or escalate alerts are consistent with the risk exposure and aligned with human expertise. Linked to (RG1), (RG6)	Bank's model / Bank	Feature combination used for few cases with justifications of the weights (divide features full list into groups for readability)
N7: Verify model control by the bank	Ensure that the bank's model does not drift over time, that there is no bias. Linked to (RG4).	Bank's model / Bank	Justify the existence and relevance of tests: Periodically draw up a list of important factors, periodic human evaluation of an alert sample

Table 6.3: Summary of supervisors' needs for model justifiability, corresponding description, model concerned and developer of justifications/explanations, and justification and explanation design ideas that emerged during the workshops.

Chapter 7

Discussion

THIS CHAPTER PRESENTS a discussion and a conclusion of the findings of this thesis. We first summarize the research contributions made in this dissertation in Section 7.1. The following sections are devoted to a discussion of our findings and future work. In Sections 7.2 and 7.3, we discuss the "explanation paradox" for decision-subjects of AI-based decisions and the human-centric avenues to improve user empowerment. In Section 7.4, we review the role of explainability to alleviate some of the regulatory tension created by black-box AI models in AML-CFT. We also highlight the relevance of the human-centric approach for implementing explainability effectively in the AML-CFT context. Finally, the discussion presents some thoughts on the lessons from the financial sector for other industries, on my experience as an interdisciplinary researcher, or on the challenge posed by Large Language Models for the explainability field.

7.1 Research contributions

In this thesis, we investigated the research question: *To what extent can AI explanations enable warranted trust and regulatory compliance in financial applications?* In Part I, we focused on the cognitive challenges for explanations to enable warranted trust, *i.e.* trust that is well-calibrated. In Part II, we explored how explanations can contribute to customer and regulator warranted trust, and enable compliance in two use cases in finance. We summarize below the research contributions presented in this dissertation.

Part I: Calibrating trust in explainable AI: common pitfalls and the promise of interactivity

Chapter 3: Trust, overtrust, distrust in explainable AI: a cognitive approach

- We provided a general vision of what and how cognitive biases affect explainability systems: with which XAI technique (*e.g.*, counterfactual explanations), user type (domain expert, AI expert or lay users) and AI-assisted task (*e.g.*, medical diagnosis).
- We highlighted how explainable AI can lead to overtrust, distrust, or how it can be misinterpreted. Some implementations of explainable

AI, however, have proven useful in correcting prior human biases in decision-making. We also emphasize that cognitive biases may affect the evaluation of explanations.

- Overall, we found that explanations usually have a tendency to increase trust, specifically for lay users, and potentially lead to unwarranted trust.
- We summarized several important factors at play in trust calibration with explainable AI systems, including user expertise, task expertise and task familiarity, estimation of the AI's confidence, explanation completeness, timing of explanations and users' motivation and individual cognitive characteristics (need for cognition, rational or intuitive decision-making style...).

Chapter 4: Towards "human-like" explanations: the promise of interactivity

- We adapted existing HCI taxonomies of interactivity to create a two-level taxonomy of interactive techniques specific to XAI, describing the interaction types and the way they support the human cognitive process of explaining: "selective", "mutable" or "dialogic".
- We analyzed the extent, nature and distribution of the interactive XAI systems included in the review.
- We offered a summary of the user-based evaluation metrics implemented in interactive XAI.
- We offered a qualitative summary of the effects of interactive explanations on several user-based evaluation metrics, finding that interactive explanations increase trust, but not necessarily overtrust, and that interactive explanations are more useful than static ones, but less easy to use and more time-consuming.

Part II: Complying with regulation using human-centric explainable AI: two case studies in finance

Chapter 5: Empowering customers of robo-advisors with explainability

- We developed a fictitious but realistic rule-based recommendation system for life insurance plans, "Robex", based on interviews with insurance supervisors and on market research.
- We created prototype explanations for Robex and redesigned them based on feedback from insurance regulators, customer protection specialists and end-users with no experience of life insurance investments.
- In our study, which involved a 2x4 between-subjects experiment with 256 participants, we found that explanations did not contribute to the legal objectives of financial regulation to empower users. Explanations did not significantly improve understanding, appropriate trust or reliance, revealing a misalignment between legal objectives and actual observed benefits of explanations.

- We highlighted how explanations still contribute to the legal objective of enhancing accountability of life insurance distributors by forcing them to provide written reasons why a given financial product is adapted to the customer's profile.

Chapter 6: Understanding the supervisors' needs for explainable AI in financial crime detection

- We described the socio-techno-legal supervision system and auditing approaches in the AML-CFT context. We reveal three main auditing approaches: global, from global to local, and from local to global. The global approach is focused on measuring the performance of the system, the global to local approach is used to sample cases where regulators discovered mistakes, and the local to global approach attempts at establishing the seriousness, and therefore the reprehensibility of the error on the whole AML system put in place by the financial institution.
- We assessed compliance obligations specific to AI-enhanced AML-CFT systems highlighting why the opacity of AI models may pose problems with regard to AML-CFT obligations.
- We formulated seven needs that supervisors have regarding model justifications and explanations. In particular, we find that supervisors primarily need to measure the performance of the AI-enhanced AML-CFT system such as gaps in detection (false negatives). However, supervisors may need contrastive AI explanations to establish the reprehensibility of sampled failure cases, to verify and challenge banks' correct understanding of the AI and to demonstrate the legitimacy of their challenger model.
- We demonstrated the complementarity of a dual HCI and legal methodology to fully understand regulatory supervisors' justification needs.
- We argued that explanations have a role of "trial evidence" to support justifications. Justifications should not only be extrinsic by referring to norms or regulations [Henin and Le Métayer, 2022], but also intrinsic by depending on faithful evidence of the system's behavior that explanations can provide.

7.2 *The potential of explanations to manipulate decision-subjects' trust*

In this dissertation, we examined whether explanations could enhance the understanding, appropriate reliance, and trust of lay users, in order to achieve the regulatory objective of user empowerment—individual autonomy, agency, free choice, informed consent—is an important objective of many legal texts imposing explanations. Specifically, we appreciated the complexity of the user empowerment problem and encountered an "explanation paradox". On the one hand, it appears logical and necessary to give individuals who are subject to an AI decision access to important information about the decision made about them. On the other hand, we revealed that explanations tend to increase unwarranted trust, and do not appear to improve significantly users' understanding of the decisions in the life-insurance context, where domain (financial) knowledge is important. Explanations play a important role in empowering end-users, while also having the potential to create inappropriate trust and reliance.

This section describes the potential for users to be manipulated through explanations. The following section will focus on the human-centric avenues that show promise for more effective explanations.

Much of the discussion below draws a comparison between meaningful consent to data practices, which has been extensively studied in the privacy literature, and meaningful consent to a decision made by an on-line AI-based recommendation system. Consent for data processing and AI recommendations share similar challenges in correcting power imbalances between data/decision subjects and data/AI operators [Acquisti et al., 2015].

7.2.1 *The Self-governance fallacy*

Our observations echo the warning of some legal scholars who have stressed that end-user meaningful consent in the digital age is a theoretical and unattainable ideal [Obar, 2020, Pasquale, 2015]. Obar [2020] characterized the situation as: "*the seemingly impossible scenario of achieving, consistently and ubiquitously, meaningful forms of consent*". This is known as the self-governance fallacy. Self-governance by end-users is an ideal that aims to empower users to understand, then consent to or decline the decisions made about them or their data. However, in the era of big data and profiling, it seems unrealistic to expect end-users to control every decision they are subject to. Pasquale [2015] argued that the "boring, time-consuming and overwhelming" nature of online consent, coupled with its mismatch with end users' real goals, who just want to use a service, make it unrealistic to expect end users to engage in "tangential" discussions about data policy (*i.e.*, information that does not have to do with the user's search). Furthermore, Morley et al. [2020] described how the self-governance approach risks creating a complex mechanism of victim-blaming in case of failure. When "empowering" an individual by providing them with choices and tools, responsibility is shifted to the individual in case something goes wrong. In the healthcare context, Mor-

ley et al. [2020] describe how an individual may be seen as a "bad actor" for failing to follow the algorithm's advice and be framed as morally responsible for his or her poor health.

7.2.2 *The dark pattern potential of explanations*

At the same time, the objective of user empowerment stems from a genuine concern that online recommendations can be harmful to end-users when the interests of online service providers and users are not aligned. Rozen et al. [2023] spoke of "dark patterns" in explainability to refer to the situation where the effect of explanations to increase trust is used to the advantage of the service provider and to the detriment of the user: *"this phenomenon of nudging users to act according to others' interest is known as "Dark Patterns" in XAI and benefits from humans' automation bias towards trusting machines* [Gray et al., 2018, Rozen et al., 2023]. In the context of data protection, Waldman [2020] argues that dark patterns exploit users' cognitive biases¹ to nudge users to cede control over their privacy. Mathur et al. [2019] define dark patterns as:

Definition

Dark patterns. *"Interface design choices that benefit an online service by coercing, steering, or deceiving users into making decisions that, if fully informed and capable of selecting alternatives, they might not make."* [Mathur et al., 2019].

¹For example, the author mentions hyperbolic discounting, a tendency to overweight immediate consequences and discount longer term ones.

Explanations can have the effect to disguise relevant or even contradictory information as evidence in favour of a product that is inappropriate for the user. In the experiment we presented in Chapter 5, participants who accepted incorrect life-insurance proposals explained in the course of dialogic explanations did not process the contradictory information presented in the explanations. Instead, the explanations had the opposing effect of reinforcing trust. Following Bösch et al. [2016]'s taxonomy of dark patterns, explanations could therefore fall into the dark pattern category of *"Hidden Legalese Stipulations"*, which consists of hiding malicious information in lengthy legal paragraphs. Alternatively, untrustworthy explanations may be included in the broader *"Sneaking"* category of Gray et al. [2018], where dark patterns are used to hide, disguise or delay information that is relevant to the user.

7.2.3 *Safeguards against user manipulation for critical online decisions*

The discussions on dark patterns or self-governance in academic literature have primarily focused on data privacy issues. In the privacy context, Waldman contends that the "predatory behavior" of online platforms is made possible because the law, *"based on the myth of rational disclosure"*, allows it [Waldman, 2020]. As a result, Waldman argues that online privacy should be better regulated by requiring large platforms to ensure the trustworthiness of their systems.

However, in the context investigated in Chapter 5, which pertains to online recommendations for life insurance contracts, recommender systems must be trustworthy by law. In finance and other highly regulated environments, regulators and internal compliance systems act as safeguards against the manipulation of user trust and dark patterns, ahead of the protection provided by user self-governance. The risk of using explanations as "dark patterns" is therefore lower for critical decisions that are subject to important regulation. In life insurance, it can be assumed that the 'fiduciary model' described by Obar [2020] is applicable. This model positions the robo-adviser company as a fiduciary, responsible for ensuring that the user's best interests are served and that relevant information is presented in an understandable manner.

However, the challenge of self-governance and consent remains prevalent in finance and other regulated industries. The legal concept of "enlightened choice" in life insurance is not solely intended for users to validate their decisions, as recommendations are expected to be reliable. Rather, it is intended to ensure that users understand the decisions they are making. This can be particularly challenging in regulated environments where there is a significant domain knowledge requirement and information asymmetry.

7.3 *Human-centric directions for improved customer empowerment*

Explanations may not always have the intended effect of improving user understanding and trust, despite regulatory expectations. Therefore, it is important to avoid the misconception that explanations are a cure-all for user empowerment and instead take a more realistic approach.

However, providing decision-subjects with relevant information on the decision remains critical and necessary, specifically for online recommendations for which human advisors are usually unavailable. The research in this dissertation shows that human-centric explainability still has an essential role to play to communicate important information to the user. Explanation interfaces may not be useful for everyone at all times, but we can optimize their design to make them "good enough", *i.e.* useful for as many users as possible, most of the time. The explanation interfaces designed in Chapter 5 offer only a few examples of the many design choices available. More research needs to be done to craft quality interactions to support customers' understanding of AI-based decisions. In what follows, I outline some promising human-centric ways of designing explanations that are worth presenting to users, and that avoid, as much as possible, the pitfalls of over-reliance and uselessness for understanding.

7.3.1 *Thinking beyond information access*

According to Obar [2020], part of the problem is that the discussion of user control and empowerment in legal and policy literature usually

ends at the point of access of information. The author states: *"once individuals have access to notice and choice manifestations, then what?"*. In legal discussion, more emphasis should be put on the *"tools for converting notice materials into meaningful consent"*. Obar also discusses *"a modified scenario where users receive summaries as opposed to details, guidance as opposed to full autonomy, support as opposed to silence"*. The turn that explainability has taken in recent years towards making explanations more visual, concise and interactive precisely aims at answering this call [Ooge, 2023].

In this thesis, we have linked legal and policy discussions on the non-expert user control problem to this current trend in explainability, which focuses on making information intelligible. The interactive, visual and dialogic explanation approaches we tested showed disappointing results in terms of end-user empowerment. However, many more explanation design strategies remain to be tested. Specifically, below I highlight that the explainability field has yet to fully exploit a wealth of research in educational psychology.

The problem is as follows: *How can explanations of online AI-based recommendations foster the empowerment of decision-subjects, specifically their understanding of decisions, and prevent user manipulation?* Below, we discuss three pathways to address the issue of client empowerment through explainability interface design:

1. Tailoring explanations to relevant user communities
2. Stimulating skepticism
3. Presenting a selected range of options
4. Fostering user engagement, curiosity and learning

7.3.2 *Tailoring explanations to relevant user communities*

In their discussion on the right to explanations for data protection, Wachter et al. [2017] highlighted that: *"What counts as a meaningful explanation for one individual or group may not be meaningful for another"*. The research community in explainability and HCI has also emphasized the importance of adapting to the needs of different user groups [Ooge, 2023, Cheng et al., 2019]. This involves striking a balance between one-size-fits-all and individualized interfaces to efficiently meet the needs of most users [Bødker, 2006]. As highlighted in [Stephanidis et al., 1999], the information society and now AI have brought us to a world where people are becoming increasingly dependent on online and AI-based services, and where AI decision subjects are not necessarily domain experts and have different skills, needs and preferences. This underlines the need for designing human-centred and high-quality technological interactions. Specifically, it requires the identification of relevant user communities, within which individuals share key characteristics influencing explanation design and have the same needs [Stephanidis et al., 1999]. The HCI discipline has a long history of "fitting" a computer artefact to a specific user group and problem setting [Avital and Te'eni, 2009]. For example, Vessey and Galletta [1991] discussed cognitive and Goodhue and

Thompson [1995] organizational task technology fit. I am hopeful that, in the near future, HCI research efforts will be able to identify the key individual cognitive factors that influence explanation effectiveness and "fit" explainability interfaces to maximize understanding among the user groups formed by these identified characteristics. To date, little is known about whether, which and how other aspects of a user's personality and profile, such as information processing styles, general intellectual ability, personal goals [Klaczynski et al., 1997], should affect the design of explanations [Naiseh et al., 2020]. In Chapter 3, we highlighted that user domain knowledge, personal goals, or need for cognition have been identified in the literature as influential in the way users process information and explanations [Klaczynski et al., 1997]. However, in Chapter 5, we did not test whether different explanation strategies could be used for different user domain knowledge. Future work could address this question.

It can be noted that adapting explanation techniques to individual profiles raises two legal challenges. First, the explainer must know something about the person receiving the explanation. This happens naturally in person-to-person communications. In online contexts, the creation of profiles, even for the sake of providing effective explanations, raises privacy concerns. Second, providing varying levels of information to different user groups may raise concerns about unequal treatment, especially if some groups receive less comprehensive information.

7.3.3 *Stimulating skepticism*

As previously mentioned, tangential explanations about the reasons for receiving the recommendation may not always be in line with the goals of the users, who are primarily interested in using the service, especially for low-stake decisions. One design approach therefore consists in forcing users to pay attention to explanations through friction. Some work has tested hiding explanations by default, or forcing users to attend explanations through friction-based interface design. For example, Buçinca et al. [2021] tested three friction-based designs: time counters, which consist in making the user wait for a certain amount of time before seeing the AI decision, on-demand buttons, which consist in displaying the explanation only on-demand of the user, and uncertainty, which consists in showing probability of the AI's prediction (e.g., "the AI is 81% confident in its suggestion").

Another possible friction-based design might be to make the warnings about the risks of the AI proposal more prominent. Buçinca et al. [2021] found that friction-based explanations reduced significantly over-reliance, at the expense of user satisfaction, however. This approach exploits users' possible suspicion that the service provider is not acting in their best interests. The lack of transparency and certainty, or the perceived risk of the AI suggestion can foster users' skepticism and critical thinking. According to Klaczynski et al. [1997], threatening problems induces more sophisticated reasoning than goal-enhancing problems. However, the effectiveness of friction-based design in improving understanding and learning has yet to be tested experimentally.

7.3.4 *Presenting a selected range of options*

Promising avenues for explainability to better align with human's cognitive architecture include Evaluative XAI [Miller, 2023], in which explanations are provided without the AI recommendations to avoid confirmation bias and clarify alternatives and trade-offs. For the same reason, some researchers have advocated presenting multiple recommendations rather than a single one. This follows the important observation that good advice does not necessarily have to be presented as a single recommendation [Miller, 2023]. While this approach may seem useful and necessary for experts such as doctors to make critical decisions, it may not be appropriate in all contexts and to current business practices, which seek to satisfy users' demand for fast, clear and therefore single advice to follow. Avoiding presenting recommendations defeats the purpose of providing a service in the first place, and providing multiple recommendations may increase the cognitive load for customers, who may not be willing to invest time and thought. As a result, offering one recommendation is often how medical (and much legal) advice is presented. Nevertheless, it seems necessary that customers invest a certain amount of time and thought if they are to make empowered decisions. We could imagine designing recommendations and their explanations in such a way that the cognitive load for customers remains low, for example by presenting a small set of relevant recommendations. For example, in situations where there are too many options to consider meaningfully, the evaluative AI framework suggests helping people narrow down the options.

7.3.5 *Fostering user engagement, curiosity and learning*

My intuition after the research in this thesis is that creating truly useful explanations requires improving user engagement or curiosity. Work on fostering motivation, curiosity, and learning in education, psychology, or HCI provides a wealth of relevant knowledge for explanation design. However, the explainability field has yet to fully tap into this research.

User engagement is related to users' motivation and goals, and to other attributes [O'Brien and Toms, 2008] such as challenge, positive affect, endurability, aesthetic and sensory appeal, attention, feedback, variety/novelty, interactivity, and perceived user control. O'Brien and Toms [2008] propose the following definition of user engagement:

Definition

User engagement. *Engagement is a category of user experience characterized by attributes of challenge, positive affect, endurability, aesthetic and sensory appeal, attention, feedback, variety/novelty, interactivity, and perceived user control.*

The explanations we designed in Chapter 5 did not improve user engagement. Future work could try to improve explanation design in the context of life-insurance in order to optimize for the above aspects. I am not aware of work in explainability that has considered all these aspects

of user engagement for explanation design.

However, interesting work has started to emerge on the tangential concept of curiosity [Danry et al., 2023, Melsión et al., 2021]. This work is rooted in educational psychology. Unlike friction-based design, supporting curiosity does not sacrifice users' satisfaction, on the contrary. According to Shin and Kim [2019], curiosity leads to a search for information which, when fulfilled, resolves the psychological discomfort of uncertainty and leads to a sense of satisfaction. In Section 2.3 of Chapter 2, we have seen that curiosity is one of the main reasons people ask for explanations. It also helps them learn and memorize better [Shin and Kim, 2019]. In the field of education, several studies have demonstrated that curiosity is a key factor in learning, creativity and well-being [von Stumm et al., 2011]. These provide promising avenues for research on explainability to promote learning through curiosity.

Definition

Curiosity. *"The desire for knowledge in the absence of extrinsic reward"* [Shin and Kim, 2019].

According to Shin and Kim [2019], curiosity is generated by the awareness of a gap in knowledge, generally aroused by stimuli [Kang et al. 2009; Markey et Loewenstein 2014]. The authors argue: *"This lack of information creates a feeling of deprivation, which naturally leads to a desire to learn."* Moreover, Shin and Kim argue that there is an optimal level of knowledge gap to arouse curiosity. Curiosity depends on how attainable the information is for them, meaning that the knowledge gap should not be too large. The feeling of having the background knowledge and ability to find an answer intensifies curiosity.

"The first step to instigate curiosity is creating an optimal knowledge gap and helping students to be aware of it. A simple way to achieve this is to introduce cognitive incongruity immediately after providing students with basic knowledge in a particular subject."

[Shin and Kim, 2019]

Asking questions to users is one way to introduce this "cognitive incongruity" and pique users' curiosity. Danry et al. [2023]'s intuition in their paper *"Don't Just Tell Me, Ask Me"* is that framing explanations as questions, rather than presenting them directly to the user, encourages people to critically evaluate explanations². They find that AI explanations framed as questions were able to significantly increase human discernment of logically flawed statements. Similarly, Melsión et al. [2021] designed "quiz" explanations by asking users—in this case children—what they thought were the most important characteristics for an AI to predict gender. The use of such gamified explanations was useful in improving understanding and learning in the domain of gender bias.

The authors' intentions in Danry et al. [2023] and Melsión et al. [2021] was to improve respectively human discernment and learning. Although the authors do not connect their research to the notion of curiosity, it

² In [Danry et al., 2023], an example of causal explanation is: *"If one person played violent video games and was aggressive, it does not follow that everyone who plays violent video games will be aggressive"*. Framed as a question, it becomes: *"If one person played violent video games and was aggressive, does it follow that...?"*

seems like designing explanations as questions corresponds to the process of stimulating curiosity described by Shin and Kim [2019]. Asking users questions makes them aware of their knowledge gaps and serves as a stimulus for curiosity. The notion of curiosity is interesting because it extends beyond simply improving critical thinking. Curiosity can prompt an active search for missing information, leading to enhanced user satisfaction and learning upon resolution. This is particularly relevant in domains with high information asymmetry, such as life insurance, where effective explanation design could capitalise on significant opportunities for learning. While fostering curiosity may seem like a worthwhile objective, it may be unattainable for some users due to time constraints and context specificities. Further research is needed to confirm or invalidate this hypothesis.

Based on Shin and Kim [2019]'s description of how to instigate curiosity, we imagined explanations designed to support it in the context of Robex, similarly as in [Danry et al., 2023]. However, due to a lack of time and resources, we did not test them. Below, we present the explanations we developed, with the hope of inspiring future researchers to empirically test similar designs.

Figure 7.1 shows the prototype interface we created. Users would first read basic information about Robex, as shown in Figure 7.1 a), to introduce basic knowledge of the Robex algorithm. Curiosity stimuli then take the form of questions as in [Danry et al., 2023]. Users would have to find the answer to two or three questions such as *"In your opinion, what feature had the most impact on the recommendation made to you?"* (Single choice question) or *"In your opinion, which of the following characteristics led Robex to make you a riskier offer?"* (Multiple choice question) as shown in Figure 7.1 b) and c). Users can click on feature cards, which turn green if it is the right answer and grey otherwise. The questions are displayed one by one to allow for progressive disclosure. [Springer and Whittaker, 2019, Panigutti et al., 2023a]. After answering a few questions, users are able to view the complete explanation, in a graphical format. The answers to the questions are saved and displayed at the top of the interface.

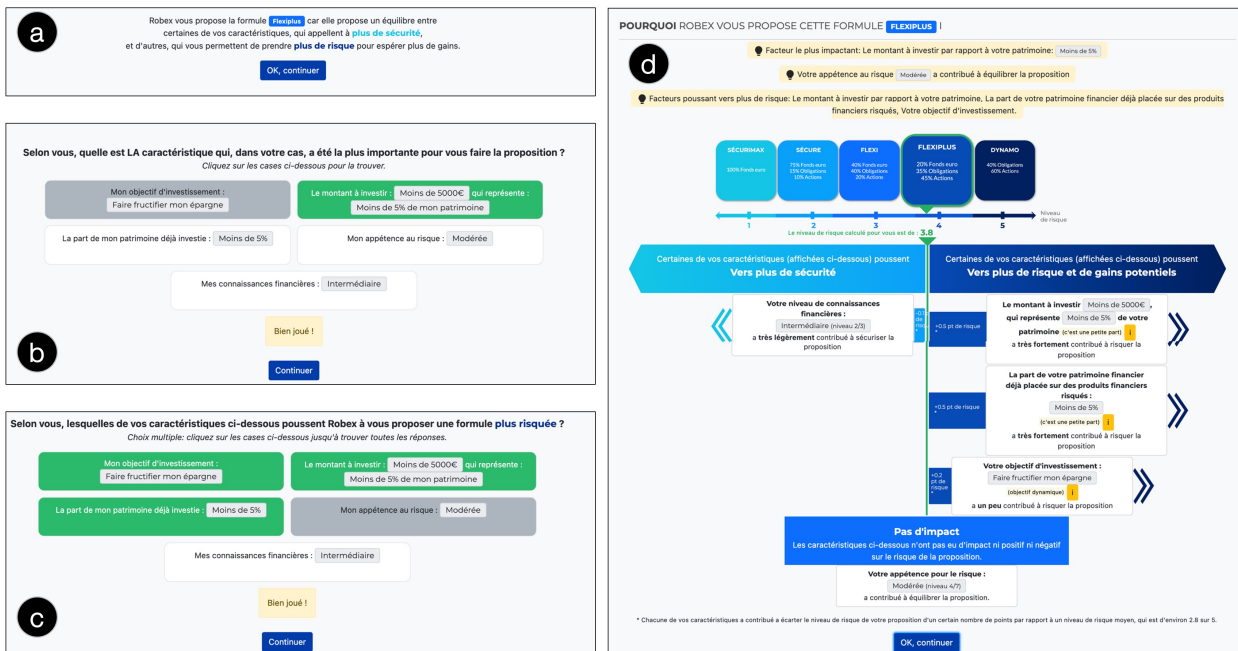


Figure 7.1: Explanation interface to engage users cognitively and stimulate their curiosity. First, a brief explanation of Robex is given: a); second, the user answers several multiple choice questions that lead them to question the impact of some features: b) and c); third, the full graphical explanation is given.

7.4 *The human-centric way forward for explainability in a highly regulated environment*

In the EU, the forthcoming AI Act will require internal compliance mechanisms and third party audits to ensure that high-risk AI systems are trustworthy. In parallel, highly regulated sectors such as finance already have in place accountability and oversight mechanisms that require all systems, including AI-based, to be trustworthy. In this context, explanations of AI systems serve to control the algorithms' outputs and verify their compliant functioning. They are directed to auditors, regulators or supervisors who are experts in the domain of application or/and machine learning. However, designing explanations for this user group also presents its own challenges, quite different from the challenges of designing explanations for lay users. On the one hand, explanations appear to ease the tension created by the use of black-box systems in highly regulated contexts. On the other hand, the limitations of current XAI techniques make them weak candidates for providing reliable and tangible evidence about machine learning's behaviour.

7.4.1 *AML-CFT illustrates the tension of using AI in a highly regulated environment*

In Chapter 6, we described the supervisory context in anti-money laundering and countering financing terrorism (AML-CFT), where the use of AI is progressing. The industry is experiencing a paradigm shift in the detection of financial crime from deterministic rule-based models, which have dominated the market for over 20 years, to probabilistic approaches using machine learning.

An increasing number of projects in banks have been utilizing AI in AML-CFT systems in recent years. AI's benefits to reduce compliance costs are beginning to materialize [Overrein, 2020], although scientific evidence that it improves the detection of money laundering and terrorism financing is still lacking. So far, financial institutions and regulators have seemed reluctant for machine learning to replace rules-based scenarios that detect known patterns of criminal activity [Blakey, 2022]. For compliance, it is important to be able to map identified AML-CFT risks to specific scenarios created in the system. Machine learning may actually be better than rules-based systems at detecting new, sophisticated, patterns of criminal activity, but the mapping exercise will be more challenging. Compliance may become less certain.

The context of AML-CFT has given us an illustration of the fundamental conflict that black box AI creates in highly regulated sectors between compliance risks and efficiency. In Chapter 6, we delved into the heart of this conflict by detailing the regulatory reasons that make AI opacity and complexity problematic. We saw that AI opacity hinders supervisors' ability to verify several key requirements of AML-CFT systems, in particular that:

1. an AML-CFT system is adapted to the specific risks of the bank's mar-

ket,

2. an AML-CFT system "carefully examines" ongoing financial operations,
3. banking analysts can justify why an AI generated alert should or should not be further examined,
4. banking operators can detect and anticipate AI failures,
5. the roles of human AML-CFT analysts and automated tools are complementary.

Additionally, the strict compliance requirements in AML-CFT create a conservative environment. To comply with AML-CFT regulation and avoid fines that can soar up to 6-7% of their turnover, banks spend dozens of billion of dollars in compliance every year and have developed costly and large-scale information systems [Farley, 2017, Goranitis and Cailali, 2023]. Updating these systems is costly and takes time [Singh et al., 2018]. Furthermore, regulators have been slow to produce guidelines on AI, enhancing the regulatory uncertainty around the use of machine learning in AML [Blakey, 2022].

All these factors heighten the tension between using AI to improve AML-CFT efficiency and compliance risks.

7.4.2 *Explainability is incomplete and uncertain*

The CJEU's *Ligue des Droits Humains* case³ requires models used to detect terrorist threats to be based on "predetermined criteria", which raises the question of whether post-hoc explainability of black box models will ever go far enough to permit the kind of verification required for critical use cases such as AML-CFT. In this section, we discuss why explainability is unlikely to fully resolve the tension between compliance and black box efficiency. In the next section, however, we argue that explainability does help to reduce this tension. We emphasize below that explainability is just one technique in the auditor's toolbox, that some XAI methods have a reliability problem, and that some explainability needs still lack computational solutions to match specific regulator needs.

³ CJUE, June 2021, 21, *Ligue des droits humains*, Case law n° C-817/19.

First, explainability only covers one aspect of the technological approaches necessary to demonstrate compliance. For example, in Chapter 6, we revealed that supervisors needed enhanced model performance metrics to compare machine learning based AML-CFT systems with pre-machine learning ones.

For the verification of the "careful examination" of ongoing financial operations by an AML-CFT system (point 2), a sampling approach is first needed to select some cases of interest with potential errors. Subsequently, explainability can be used to determine whether the algorithm's examination of specific cases indeed contains methodological shortcomings. Challenger models, such as those used by the ACPR in France, also seem particularly relevant and necessary to challenge point 2.

Similarly, explainability seems insufficient to fully demonstrate that banking operators can detect and anticipate AI failures (point 4). What

seems necessary in this case is a demonstration of a high-quality model governance, which goes beyond the scope of explainability.

Second, current XAI techniques have a reliability problem. Currently, there is a lack of assurance that the concept of explainability is one hundred percent truthful [Bilodeau et al., 2023, Kindermans et al., 2017]. As presented in Section 6.5.2 of Chapter 6, feature-based explanation techniques are based on correlations between features, not on causation [Hammon et al., 2022, 2020, Rouvroy, 2013], making it hard for regulators to rely on explanations as "faithful" and factual evidence for justifications. Explanations can also be manipulated in black-box audit settings so as to hide potential biases in a model, as demonstrated by Zhou and Joachims [2023]. In response, Jeannette Wing advocates for the use of formal methods to address the probabilistic nature of machine learning and the role of data in training with a deterministic tool [Wing, 2021]. Formal verifications, she argues, are needed complements to fairness, robustness, accountability, and explainability in order to achieve trustworthy AI. Additionally, some work on the causability of explanations such as [Holzinger et al., 2020] are promising to address some of the inherent flaws of current explainability methods [Confalonieri et al., 2021].

Third, our findings in Chapter 6 pointed to supervisors' explainability need to establish the level of reprehensibility of sampled failure cases (point 2): "Was the failure an isolated incident or does it reveal a more serious systemic problem?" However, Zhou and Joachims [2023] argue that current explanations do not provide answers to questions like: "what factors caused the model to predict X instead of Y?", although this is precisely what supervisors are looking for in AML-CFT: "what factors caused the model not to produce an alert for this case (*instead of flagging the case*)? Computational solutions to provide such explanations are indeed lacking in the explainability literature [Miller, 2021]. Future research in explainability could investigate if contrastive explanation models such as in [Miller, 2021] could provide solutions to this problem.

7.4.3 *Human-centric explainability alleviates some of the regulatory tension of black-box AI*

Nevertheless, explainability can ease some of the tension for regulatory compliance caused by AI opacity. Explanations help to determine whether a decision was made in accordance with procedural and substantive standards, which is the first aspect of accountability as defined in [Doshi-Velez and Kortz, 2017]. Explainability also contributes to accountability by providing evidence to support the justifications made by the regulated entity [Felici et al., 2013]. The evidence may be imperfect due to the reliability problem highlighted above, but at least some evidence will be present.

The list below⁴ gives our assessment on the level of contribution of XAI to the five regulatory requirements listed above. The points 1' and

⁴This argumentation format is inspired from [Miller, 2023].

2' describe some of the technical functions that XAI methods can perform, contributing to demonstrate compliance respectively to the points 1 and 2 in the list above. Furthermore, explainability may contribute to answer the regulatory issues presented in points 3, 4, and 5. However, the predominant human element in these contexts of XAI use makes the adoption of a human-centric approach to explainability design particularly critical.

- 1'. **can** reveal if certain characteristics of a bank's clientele and risk profiles are duly taken into account, through global XAI,
- 2'. **can** reveal if the algorithm's "examination" of operations contains methodological errors, and how it adjusts to new information, using local and global XAI methods,
- 3'. **may** enable an analyst to understand an alert and produce quality reports using local XAI, provided that human cognitive biases and human factors are carefully accounted for,
- 4'. **may** help banking operators to demonstrate control over their AI system,
- 5'. **may** allow better coordination between machine and human analysts and more timely processing of alerts.

In a qualitative enquiry with similar AML-CFT scenarios as we used in 6, Gerlings and Constantiou [2022] found that contextual explanations were much needed to enable banking investigators to understand an alert produced by machine learning in a timely manner. Explanations can therefore contribute to points 3' and 5'. However, they also noted the risk of investigators being influenced by an alert's risk score⁵ and losing time trying to understand it. The authors suggest removing such scores altogether or providing more context-relatable explanations to point investigators to the issues with an alert.

⁵ "If the score is low, effort is low and vice versa." [Gerlings and Constantiou, 2022]

In Chapter 6, we found that supervisors need explainability to verify that the AI's criteria for escalating or closing alerts are consistent with human expertise. However, more research in HCI is needed to develop useful explainability interfaces for supervisors to verify the bank's control over its model and for banking practitioners to detect and anticipate errors (points 3 and 4).

We also described in Chapter 6 the complexity of the socio-techno-legal context of AML-CFT supervision. We found that supervisors mainly had legal backgrounds, with few investigators having AI development knowledge. The holistic perspective provided by human-centric approaches will be particularly important to design explanations for supervisors that take into account these social factors.

In the complex and high-dimensional context of AML-CFT, the human-centric approach strikes me as particularly necessary for effective explainability implementation. It allowed us to uncover the need for contrastive explanations among supervisors, which can pave the way for adapted computational XAI solutions that respond to this need. The human-centric approach will also be necessary to ensure that banking analysts are not biased by explanations, that development teams feel accountable

for their models, that explanations improve human-machine collaboration and that supervisors with high domain expertise but little computer science skills have the means to challenge bank's implementation of explainability. Explainability should not be seen as an off-the-shelf solution, but as one tool among many in a complex socio-techno-legal context.

7.5 *Peripheral observations*

7.5.1 *Why the financial sector is interesting for other highly-regulated industries*

As our discussion is based on the finance case studies presented in Part II, we highlight below two reasons why the results we presented in the financial sector can provide insights for other highly regulated industries.

First, the risk-based approach used in AML-CFT is a common regulatory approach, specifically in law pursuing "crime-fighting and public safety objectives" [Black, 2001]. The lessons we learned in the area of AML-CFT therefore resonate in these other risk-based fields. The approach is generally presented as virtuous because of its proportionality and cost-effectiveness [OECD, 2021b]. It is adopted, for example, in the recent Digital Services Act to prevent the systemic risks posed by AI-based information platforms [European Parliament and Council, 2022] or in the Draft Regulation on the Dissemination of Terrorist Content Online [European Parliament and Council, 2021, Maxwell, 2021]. Maxwell also notes some downsides to this approach. One of its peculiarities is that it shifts the burden of attaining public interest objectives through 'appropriate' means to private actors, which are not as directly accountable as are public authorities for respect of fundamental rights. Where regulatory compliance is measured in part by the quantity of resources devoted to a detection or enforcement task, this can incentivize companies "to do too much, rather than too little, to satisfy the law's crime-fighting objectives, a phenomenon known as gold-plating." Gold plating can in turn create risks for fundamental rights by going beyond what is strictly necessary and proportionate.

Second, the digital developments we have seen in this thesis in the financial sector with the emergence of online robo-advisors and supervisory technology tools for AML-CFT, such as "LUCIA", are likely to be adopted in other areas of the regulated digital economy.

Current global efforts to regulate technology pose unprecedented challenges for regulators and create a demand for new regulatory technology. The financial sector has been at the forefront of the development of tools to support compliance and reporting. The rapid development of FinTech in the aftermath of the 2008 crisis, together with the burdensome compliance measures in the financial sector, have necessitated a corresponding evolution of regulatory tools. RegTech addresses this demand by providing software tools that support regulatory compliance. It has been instrumental in catalysing innovation and allowing digital companies to

navigate in the complex financial compliance landscape [Paul Fehlinger, 2023].

Definition

RegTech. *“Any use of technology to match structured and unstructured data to information taxonomies or decision rules that are meaningful to both regulators and the firms they regulate, in order to automate compliance or oversight processes.” [Emmanuel Schizas et al., 2019].*

In summary, the risk-based approach to regulation and the pioneering regulatory developments in the financial sector suggest that lessons learned in this area could be instructive for other highly regulated sectors.

7.5.2 *Principles for dealing with interdisciplinarity*

This thesis has underlined the need for interdisciplinarity in XAI research. Interdisciplinary, however, is challenging. Acquiring adequate proficiency in a single field demands extensive practice, making it particularly challenging to attain expertise in multiple areas. For a novice researcher, not fully established in any research domain, interdisciplinarity can therefore seem like an impossible endeavour, running the risk of making no contribution anywhere and tackling subjects only superficially. I was confronted with this problem throughout my thesis. Moreover, building on different fields can make it particularly difficult to have relevant experts review the scientific value of interdisciplinary contributions, and ensure their quality. Below, I highlight three principles that I believe are important, although very simple, and not specifically original, to address the challenges of interdisciplinarity.

1. Clarifying one’s roots. First, interdisciplinary authors should clarify the disciplinary origin(s) of the methodologies used. Interdisciplinary contributions sometimes lack a clear indication of the field or literature they draw upon. This can make it difficult to evaluate their scientific value and can contribute to the undermining of interdisciplinary research. However, if relationships to disciplines are clearly specified, relevant reviewers can be called upon to verify research quality. Furthermore, I have realized in my research the importance of borrowing established methods from academic disciplines (in my case mainly from HCI). Using established methods allows for capitalising on decades of evolution in a field towards scientific value, and provides the opportunity to demonstrate rigour, transparency, and accuracy in their application. It also allows peer reviewers to assess the quality of the implementation of the methods.

2. Establishing a shared vocabulary and knowledge base. Second, interdisciplinary fields have to establish a shared knowledge base and vocabulary among researchers from diverse backgrounds. Dealing with various terms, multiple definitions for the same concept, and diverse backgrounds is a well-known challenge faced by interdisciplinary research communities. In explainability, several researchers have called

for more unity and consensus in the vocabulary used [Doshi-Velez and Kortz, 2017, Markus et al., 2021]. However, the idea is difficult to put into practice, as in 2023, divergences still exist on the definition of explainability. Notable efforts to map the landscape of interdisciplinary research on AI ethics, transparency or fairness are presented in [Jobin et al., 2019, Abdul et al., 2018]. They provide useful insights on the complex sub-communities that form the interdisciplinary research field on AI. Other useful initiatives are workshops and courses provided in interdisciplinary conferences. They contribute to give all authors and reviewers a minimum understanding of the different approaches and theories relevant to the field.

3. Embracing historical research. Third, emerging interdisciplinary movements and fields may sometimes lack sufficient connection to their historical research roots. This is related to my first two points. It can be hard to realize that some research has already been done on a topic if the terminology employed was not exactly the same. Specifically, human-AI interaction research should better embrace past research in HCI and psychology to build upon it. For instance, the field of psychology and visualisation has generated a considerable amount of literature on how to communicate information effectively to individuals, without necessarily referring to the concept of explanation. Yet, explainability would have much to gain from these findings. If explainability does not recognise its links with these disciplines, it not only misses the opportunity to capitalise on relevant knowledge but also runs the risk of replacing such knowledge with more recent studies that may not be based on as well-established methodologies.

7.5.3 On explainability for LLMs

Significant developments have occurred in the field of explainability in 2023, driven by research on LLMs. The unprecedented size of large language models (LLMs)⁶, their dependence on context thanks to attention mechanisms [Vaswani et al., 2017], and their capture of the intricate nuances of language have fascinated many researchers. LLMs also present new risks [Geburu et al., 2021]. Specifically, they suffer from "hallucination", *i.e.* generating inaccurate, non-factual content [Yao et al., 2023a] and their mode of interaction with people through dialogue, as we have seen in Chapter 5, makes them particularly prone to cheat users and persuade them of false claims [Rozen et al., 2023]. As Bubeck et al. [2023] puts it: "*[GPT4] is remarkably good at generating reasonable and coherent explanations, even when the output is nonsensical or wrong*". This has driven many scholars to attempt to better understand the underlying mechanisms of LLMs. The last couple of years have therefore seen interesting developments in the field of explainability, from an observation-based, "natural science" approach to a more promising, mechanistic and engineering approach.

One of the specificities of LLMs like GPT-4 is that it can give you an explanation of its answers if you ask it to. A lot of the efforts to better understand large language models have therefore focused on design-

⁶The phenomenon of emergence in LLMs refers to the abilities that are not present in smaller language models but appear when scaling up models [Wei et al., 2023].

ing inputs or "prompts" that elicit explanations. For example, Chain-of-Thought prompting (CoT) consists in eliciting intermediate reasoning steps in the LLM's output [Wei et al., 2023]. Many more strategies in that vein have been developed to improve end task performance, for example few-shot prompting which consists in giving an example of the expected result in the prompt [Brown et al., 2020] or ReAct [Yao et al., 2023b] which instructs the model to perform specific actions such as searching an external information source. Bubeck et al. [2023] also tested GPT-4's explainability abilities by asking it to provide explanations for its answers. They examined its *output consistency*, *i.e.* whether the explanation given by GPT-4 is consistent with its output, and its *process consistency*, *i.e.* whether the explanation gives us the ability to simulate GPT-4's predictions in different similar contexts. They found that GPT-4 was particularly output-consistent, even when providing an explanation for a wrong answer, but not reliably process-consistent, especially for tasks that are not inherently explainable, such as arbitrary ones.

Although these strategies are called "prompt engineering", they deviate from the idea of understanding the algorithms' internal components through formal engineering and mathematical methods. In this sense, they are more closely aligned with natural science approaches.

Most of the above-mentioned approaches rely on inference, observations, and more specifically on the language models' outputs. However, since LLMs' outputs are unreliable [Yao et al., 2023a] there is no guarantee that prompting strategies will make their answers and explanations more accurate. Turpin et al. [2023] recently demonstrated that Chain-of-Thoughts prompting can fail and generate false reasons for the chatbots' answers in the step-by-step reasoning. Moreover, the consistency of GPT-4's output presents a significant issue. If the responses are inaccurate, the corresponding explanations will align with them and convince users of erroneous assertions [Bubeck et al., 2023].

The classical black-box approaches to explainability provided by methods like SHAP, counterfactual and other model-agnostic techniques have also been tested on LLMs. Martens et al. [2023] have even taken advantage of the LLMs to provide "SHAPstories" and "CFstories", narratives generated from the results provided by these techniques. They show that these narratives are more convincing for human users, providing useful tools to generate explanations to a general audience and nonspecialists, they argue. Yet, these approaches have the limitations we know of classical explainability methods, specifically lack of causability, in addition to the limitations of prompt-based explanation methods such as non-robustness due to high sensitivity to prompt details combined with output consistency problems and persuasiveness.

Some recent research has introduced promising results to "mechanistic" explainability, *i.e.* explain models' internal mechanisms and components. Such advances have been made possible by experimenting with small models. Early attempts at understanding LLMs and deep learning models have focused on trying to find what best activates individual neurons [Nguyen et al., 2016, Carter et al., 2019]. However, the activation of a single neuron can take many different meanings in different

contexts, which makes it impossible to interpret neural networks on this basis. This is what Anthropic [2023] call the *polysemanticity* of individual neurons. This can be due to the *superposition* phenomenon by which "a neural network represents more independent "features" of the data than it has neurons by assigning each feature its own linear combination of neurons." [Anthropic, 2023]. However, recent research by Anthropic has shown that mechanistic explanations are possible on small models at the feature scale, which is much more appropriate than the scale of a single neuron. By analyzing patterns (linear combinations) of neuron activations, they provide a promising path to breaking down the complexities of neural networks into parts we can understand. For the first time, it feels like the mechanistic approach could be surmountable, and explainability could be achieved through a formal rather than purely inference-based methods. These findings have yet to be replicated on larger, "frontier" models, however.

7.6 *General conclusion*

The first part of this thesis examined the impact of explainability on appropriate trust through two detailed scoping reviews focusing respectively on cognitive biases and interactive explainability. We established that explanations have the potential to manipulate trust, by triggering cognitive mechanisms that lead to overtrust, distrust or misusing algorithmic explanations and predictions. We documented some factors that play an important role in the trust calibration process with AI systems, namely users' prior beliefs and knowledge, and the completeness, framing and the timing of the explanation. Interactivity has recently been advocated by some scholars as a possible way of better aligning explainability interfaces with the human cognitive processes of explanation. Therefore analysed the different types of interaction found in the literature on explainability and summarised the effects of interactivity on explainability. Currently, interactive explanations do not appear to increase misplaced trust in AI systems. However, there is a scarcity of relevant controlled experiments to confidently confirm or refute this.

In the second part of the dissertation, we explored the role of explanations for appropriate trust, which is critical for AI compliance in two case studies in finance.

In the domain of life-insurance distribution, we came across an "*explanation paradox*". Explanations are intended to empower users by providing them with important domain knowledge to enable them to make free, informed choices. However, explanations also have the potential to increase unwarranted trust and make users more vulnerable to untrustworthy recommendations. In these circumstances, it appears challenging, if not unattainable, for explanations to meet regulatory expectations of ensuring meaningful consent from each and every individual. As highlighted in Section 7.3, explanations should not be seen as a silver bullet for empowering customers. However, future work could explore how to develop "better than nothing" explanations that work fairly well for most people. Promising work in explanation design is moving in this direction by studying how explanations can be tailored to relevant client groups, how friction-based interface design can be used, and designs that support curiosity and learning.

In the domain of anti-money laundering and countering terrorism financing, we have discovered that explanations are necessary to enable regulatory supervisors to trust (or not) AML-CFT systems operated by financial institutions. Explanations can provide evidence on AI systems' behaviour. Such factual information supports the provision of justifications—*i.e.* demonstrations of compliance—by regulated entities. For example, explainability will be necessary to verify the alignment of machine and human criteria for flagging money laundering cases, and less clearly to verify the appropriate prevention of potential AI failures. We also established the need of supervisors for contrastive explanations that help to determine the level of reprehensibility of sampled failure cases: "Was the failure an isolated incident or does it reveal a more serious systemic problem?". However, computational solutions remain to

be developed to address this need. Additionally, we noted that current explainability methods have reliability issues that need to be resolved. We argued that taking a human-centric approach is crucial in mitigating the regulatory tensions caused by the use of opaque machine learning in the complex socio-techno-legal environment of highly regulated sectors such as AML-CFT.

Below are some short recommendations for future research and policy. These recommendations reflect my subjective interpretation of the results of my thesis.

Recommendation 1. Examine the needs of online robo-advisor clients in more detail. This will help to better align them with regulatory objectives. Further qualitative research should delve into the needs of different types of robo-advisor clients in light of the regulatory objectives they are intended to fulfill.

Recommendation 2. Determine whether friction-based explainability design can improve user understanding and critical thinking, even marginally. Some work has started to investigate how to force users to pay attention to explanations through "friction" [Buçinca et al., 2021, Naiseh et al., 2021a]. Further work could explore the effect of explanations that use prominent risk warnings or that only appear if requested, on user understanding of an AI recommendation

Recommendation 3. Examine the impact of question-driven explainability design to optimize curiosity and learning. Absence of domain knowledge can create obstacles to users' effective understanding of AI recommendations. Explainability should be viewed as an opportunity to educate consumers on basic domain knowledge. Formulating explanations as questions [Danry et al., 2023, Melsión et al., 2021] can be useful in sparking consumer curiosity and learning. Research in educational psychology should be leveraged to make sure explanations can foster curiosity.

Recommendation 4. Take a human-centric approach for explainability use in AML-CFT and other complex socio-techno-legal environments. Explainability should not be viewed as a ready-made solution, but rather as one tool among many in a complex socio-techno-legal context. Therefore, we emphasise the importance of designing explainability with a human-centric approach, taking into account the diverse backgrounds, needs, feelings of accountability, and cognitive biases of different stakeholders. This approach can be complemented by legal analyses to better understand regulatory requirements, which go hand in hand with the needs of supervisors.

Recommendation 5. Develop and design contrastive explanations to help supervisors gauge the level of reprehensibility of failure cases. The aim of this exploration would be to answer the supervisor's question: 'Was the failure an isolated incident or does it reveal a more serious systemic problem?' At present, XAI techniques provide inadequate solutions to this issue.

Recommendation 6. Elaborate tests to verify the correct human and AI alignment of decision criteria and prevention of failures. As we have seen, machine learning in highly-regulated tasks such as AML-CFT must permit regulated entities and supervisors to verify alignment of the system with human-defined decision criteria. Current ex-post XAI techniques do not permit this yet, but XAI developments are quickly advancing so that this alignment can be verified in the near future.

Appendix

A1. List of cognitive patterns when interpreting explainable AI

Table A.1: List of cognitive patterns identified in the corpus created in Chapter 3 that may lead to reasoning errors when using explainable AI systems.

Cognitive pattern	Definition	Ref. in the corpus
<i>Ambiguity aversion</i>	"The tendency to prefer known risks over unknown risks" [Kliegr et al., 2021]	[Kliegr et al., 2021]
<i>Anthropomorphism</i>	People tend to attribute human traits to machines and therefore expect AI explanations to use the same conceptual framework used to explain human behaviors.	[Miller, 2019, Weld and Bansal, 2018]
<i>Attention to aesthetics</i>	Human judgment ratings of explanations are biased toward visual appearance.	[Mohseni et al., 2021a]
<i>Attention to abnormality</i>	"People mostly ask for explanations of events that they find unusual or abnormal" [Miller, 2019]	[Miller, 2019, Weld and Bansal, 2018]
<i>Attention to confidence levels</i>	People need confidence levels to make better use of ML-assisted decision-making systems. "Prospect Theory suggests that uncertainty (or risk) is not considered independently but together with the expected outcome" [Bhatt et al., 2020]	[Bhatt et al., 2021, Miller, 2019]
<i>Attention to demographic features</i>	Tendency to fixate on demographic features in explanations such as age and race	[Liu et al., 2021]
<i>Attention to False Negatives rather than to False Positives</i>	"Users pay less attention to FP explanation errors and in turn, are more critical for FN explanation errors". [Mohseni et al., 2021a]	[Mohseni et al., 2021a]
<i>Attention to foil</i>	"Explanations are sought in response to particular counterfactual cases, which are termed foils. That is, people do not ask why event P happened, but rather why event P happened instead of some event Q." [Miller, 2019]	[Miller, 2019, Weld and Bansal, 2018, Woodcock et al., 2021]
<i>Attention to intentionality and responsibility</i>	People tend to focus on intentional actions rather than non-intentional ones to select an event as a cause in a causal chain. Similarly, "an event considered more responsible for an outcome is likely to be judged as a better explanation than other causes."	[Miller, 2019, Weld and Bansal, 2018]
<i>Attention to necessity, sufficiency and robustness</i>	Events that are necessary, sufficient and robust to some changes are more likely to be selected as a cause.	[Miller, 2019]
<i>Automation bias / automation overreliance</i>	The tendency to over rely on machine's predictions.	[Bansal et al., 2021, Bussone et al., 2015, Danry et al., 2020, Liu et al., 2021, Naiseh et al., 2021b]
<i>Availability bias</i>	The tendency to believe that examples and events that easily come to mind are more representative than is actually the case.	[Kliegr et al., 2021, Wang et al., 2019a, Zytek et al., 2021]
<i>Averaging bias</i>	"Using the average of probabilities of two events for the estimation of the probability of a conjunction of the two events". [Kliegr et al., 2021]	[Kliegr et al., 2021]
<i>Backfire effect</i>	"Corrections of misperceptions may enhance people's false beliefs". [Nyhan and Reifler, 2010]	[Lai and Tan, 2019]
<i>Base-rate neglect</i>	"The tendency to underweight evidence provided by base rates". [Kliegr et al., 2021]	[Kliegr et al., 2021]
<i>Change blindness Choice overload</i>	"Humans inability to notice all of the changes in a presented medium". [Simons, 2000] The difficulty to make a choice when facing many choices for people of the type "mazzimizer". As a consequence, they are less committed to their choices, display lower satisfaction with their choices.	[Sokol and Flach, 2020] [Coba et al., 2019]
<i>Cognitive dissonance</i>	The tendency to agree with the AI's suggestions, while being aware to have a different opinion.	[Danry et al., 2020]
<i>Completeness bias</i>	Longer explanations tend to lead more to overreliance than shorter ones.	[Bussone et al., 2015, Fürnkranz et al., 2020, Kulesza et al., 2015, Lai and Tan, 2019, Szymanski et al., 2021]

Cognitive pattern	Definition	Ref. in the corpus
<i>Confirmation bias and hindsight bias</i>	"The tendency to seek supporting evidence for one's current hypothesis". [Kliegr et al., 2021]	[Bayer et al., 2021, Kliegr et al., 2021, Bussoni et al., 2015, Naiseh et al., 2021b, Szymanski et al., 2021, Wang et al., 2019a]
<i>Confusion of the inverse</i>	"The mistake of confusing the confidence of an implication $A \rightarrow B$ with its inverse $B \rightarrow A$." [Kliegr et al., 2021]	[Kliegr et al., 2021]
<i>Conjunction fallacy</i>	Estimating the conjunction of two statements to be more probable than one of the two statements.	[Fürnkranz et al., 2020, Kliegr et al., 2021, Weld and Bansal, 2018]
<i>Default or Status quo bias</i>	"The tendency to favor the default option and thus the proposed suggestion". [Bayer et al., 2021]	[Bayer et al., 2021]
<i>Disjunction fallacy</i>	"Judging the probability of an event as higher than the probability of a union of the event with another event". [Kliegr et al., 2021]	[Kliegr et al., 2021]
<i>Disregard of evidence</i>	Tendency to believe persuasive claims unsupported by evidence.	[Danry et al., 2020]
<i>Escalation of commitment</i>	"People stick to a choice they made despite understanding the logical implication that doing so might lead to undesirable consequences" [Bayer et al., 2021]	[Bayer et al., 2021]
<i>Familiarity bias</i>	"Unfamiliar information might induce a reinforcement effect that causes users to avoid interacting with various content". [Szymanski et al., 2021]	[Szymanski et al., 2021]
<i>Framing bias</i>	People decide on options based on whether they are presented with positive or negative connotations or whether they are presented after or before the AI recommendation.	[Bansal et al., 2021, Bhatt et al., 2021, Kim and Song, 2020, Kliegr et al., 2021]
<i>Illusion of Explanatory Depth</i>	People think they have a much deeper understanding of how complex concepts work than they actually do.	[Chromik et al., 2021, Kaur et al., 2020, Naiseh et al., 2021b]
<i>Illusion of validity</i>	"Unjustified sense of confidence and hence failure when evaluating different possibilities" [Simkute et al., 2020]	[Simkute et al., 2020]
<i>Illusory superiority</i>	"Users with the highest need for advice may be the least likely to defer judgment." Also known as the Dunning-Kruger effect [Schaffer et al., 2019].	
<i>Inherence bias</i>	"Humans tend to construct explanations based on accessible information about the inherent properties of a particular phenomenon instead of inaccessible information about extrinsic factors". [Bekele et al., 2018]	[Bekele et al., 2018, Miller, 2019]
<i>Information overload</i>	"Providing too much information at once can result in reduced accuracy" [Simkute et al., 2020]	[Abdul et al., 2020, Naiseh et al., 2021b, Simkute et al., 2020, Zyteck et al., 2021]
<i>Insensitivity to sample size</i>	When both confidence and support are stated, confidence scores positively affects plausibility and support is largely ignored.	[Fürnkranz et al., 2020, Kliegr et al., 2021]
<i>Insensitivity to sample variance</i>	"Users are primarily guided by the mean and the number of ratings, and to lesser degree by the variance and origin of a rating" [Coba et al., 2019]	[Coba et al., 2019]
<i>Mere exposure effect</i>	The increase of trust in an AI suggestion following the mere exposure of an explanation.	[Eiband et al., 2019, Kliegr et al., 2021, Lai and Tan, 2019]
<i>Misunderstanding of Boolean logic</i>	"People interpret "AND" differently than logical conjunction", the TRUE and FALSE conditions are perceived as non-intuitive. [Kliegr et al., 2021]	[Kliegr et al., 2021, Fürnkranz et al., 2020]
<i>Misunderstanding of confidence scores</i>	Not understanding what the confidence scores refer to.	[Bussoni et al., 2015]
<i>Narration bias (linked to over-generalization)</i>	Tendency to interpret information as being part of a larger story and to assume causal relations in the events of that story.	[Andrienko et al., 2022, Atrey et al., 2020, Kaur et al., 2020, Zyteck et al., 2021]
<i>Negativity bias</i>	Users pay more attention to negative features in the AI or the AI explanations which may lead to eroding trust and pay more attention to negative outcomes.	[Branley-Bell et al., 2020, Kliegr et al., 2021, Nourani et al., 2021, Shimojo et al., 2020, Zyteck et al., 2021]
<i>Perceived goal impediment</i>	"People in highly critical decision-making environments are likely to be in a serious-minded state, where additional information might be prone to being perceived as a goal impediment".	[Naiseh et al., 2021b]
<i>Pre-use algorithmic optimism</i>	Before using the XAI system, users had positive inferences about algorithmic capability, which disappeared after using it.	[Springer and Whittaker, 2019]
<i>Preference for broad explanations</i>	People prefer broad explanations, that explain more observations.	[Miller, 2019]
<i>Preference for more complete explanations</i>	People tend to prefer complete explanations over sound ones. Complete explanations help them form better models.	[Kulesza et al., 2013]
<i>Preference for simple explanations</i>	People prefer simple explanations to complex ones.	[Abdul et al., 2020, Miller, 2019, Shimojo et al., 2020, Zyteck et al., 2021]
<i>Preference for usability vs. performance</i>	User performance and preference on proxy tasks may not accurately predict their performance and preference on the actual decision-making tasks where their cognitive focus is elsewhere, and they can choose whether and how much to attend to the AI.	[Bućina et al., 2020, Liu et al., 2021, Szymanski et al., 2021]
<i>Primacy effect or Anchoring bias</i>	People quickly form opinions about something based on the first information we receive about it.	[Kliegr et al., 2021, Naiseh et al., 2021b, Nourani et al., 2021, Wang et al., 2019a]
<i>Recognition bias</i>	Recognizing information makes the user more likely to trust the explanation.	[Fürnkranz et al., 2020, Kliegr et al., 2021, Szymanski et al., 2021, Woodcock et al., 2021]
<i>Redundancy aversion</i>	Redundant information is another cause of skipping explanations, making users lose trust in the explanations.	[Naiseh et al., 2021b]
<i>Reinforcement effect or Reiteration effect</i>	The increase of trust following repetition.	[Kliegr et al., 2021]
<i>Representativeness bias</i>	The similarity of objects or events makes people disregard the probability of an outcome.	[Fürnkranz et al., 2020, Kaur et al., 2020, Kliegr et al., 2021, Wang et al., 2019a, Zyteck et al., 2021]
<i>Unit bias</i>	"The tendency to give a similar weight to each unit rather than weigh it according to its size". [Kliegr et al., 2021]	[Kliegr et al., 2021]
<i>Weak evidence effect</i>	"Weak argument in favor of a statement can lead to decreased believability of the statement". ([Kliegr et al., 2021]	[Kliegr et al., 2021, Fürnkranz et al., 2020]

B1. Co-design Study Questionnaire

Figure B.1: The following figure presents the questions used in the co-design interviews conducted in Chapter 5.

Each interview included the following steps:

1. **Preliminary questions:** End-user participants are asked questions on their experience with life-insurance and robo-advisors, and on their explanations needs.
- 1'. **Preliminary questions:** Regulator participants are asked questions about explanations' role for customers and customer protection in life-insurance
2. **Testing the interface:** Participants are asked to use Robex from the profiling questionnaire up to the recommendation and explanation. They are also asked to think aloud. Regulators are prompted to use Robex with several different imaginary user profiles.
3. **Feedback:** Participants are asked for feedback about their overall experience using Robex.

Below are the questions asked to participants. The questions have been adapted slightly depending on whether they were asked to regulators or end-users. Questions for regulators are shown in the **blue boxes**, those for non-expert participants in the **red boxes**. The **purple boxes** indicate that there was no difference between the questions asked to regulators and end-users for the phase in question.

Phase 1: Preliminary questions (Regulators)

1. How important are explanations for users in life insurance? What type of explanations should be provided?
2. How good are the explanations offered by robo-advisors?
3. How can we reach people with no financial knowledge?
4. What do you think potential subscribers need to make an informed decision?

Phase 1: Preliminary questions (End-users)

1. Do you have any experience of using a robo-advisor or life insurance?
2. What is your level of familiarity with financial investment?
3. What kind of explanations would you like to receive about an online financial recommendation?

Phase 2: Testing the interface

1. Do you agree with the proposal?
2. What would you have suggested?
3. Do you agree with the explanations?
4. Test another profile

Phase 3: Feedback

1. What is your experience / opinion of the system? Do you think these explanations could help users?
2. What do you think of the proposed explanations? Are there any limitations, other needs?
3. What user characteristic would it be interesting to change in the explanations?

B2. The Robex recommendation system

We describe below the simple, rule-based scoring algorithm for Robex.

o, a_s, c, a_p, k represent the dimensional risk scores obtained by a user after responding to a profiling questionnaire. o represents the user's financial objective, a_s her assets, c her asset composition, a_p her risk appetite and k her knowledge in finance. Dimensional risk score values were calibrated through multiple discussions and tests with regulators. rs is the total risk score, the sum of the dimensional risk scores.

$reco$ is Robex's recommendation. 1 is the least risky and 5 is the most risky.

Ensure:

```

 $\mathbb{R} \leftarrow o, a_s, c, a_p, k, rs$ 
 $\mathbb{Z} \leftarrow reco$  with  $1 \leq reco \leq 5$ 
 $0 \leq o \leq 3$ 
 $-2 \leq a_s \leq 4$ 
 $-9 \leq c \leq 1$  with  $c = f(a_s)$ 
 $0 \leq a_p \leq 7$ 
 $0 \leq k \leq 5$   $rs = o + a_s + c + a_p + k$ 
if  $rs < 6$  then
   $reco \leftarrow 1$ 
else if  $rs < 10$  then
   $reco \leftarrow 2$ 
else if  $rs < 15$  then
   $reco \leftarrow 3$ 
else if  $rs < 19$  then
   $reco \leftarrow 4$ 
else
   $reco \leftarrow 5$ 
end if
  Additionally,  $\triangleright$  Safety measures where added for specific user
  answers.
if  $o = 0$  then
   $reco \leftarrow \min(reco, 2)$ 
end if
  and
if  $a_s = 0$  then
   $reco \leftarrow 1$ 
end if

```

Algorithm 1: The Robex rule-based algorithm

The biased Robex algorithm works like this: the total risk score rs that is obtained by a user is artificially reduced or increased by about 10 points, which amounts to the total false risk score frs . The following algorithm calculates false dimensional risk scores o, a_s, c, a_p, k that together sum up to the total false risk score frs .

Ensure:

```

 $\mathbb{R} \leftarrow \text{frs}$ 
 $\text{frs} \leftarrow \text{rs}$ 
 $\mathbb{R}^5 \leftarrow w$  is the array of  $o, a_s, c, a_p, k$  values sorted in descending order
 $\mathbb{R}^5 \leftarrow \text{MAX}$  is the array of maximum values for  $o, a_s, c, a_p, k$ 
 $\mathbb{R}^5 \leftarrow \text{MIN}$  is the array of minimum values for  $o, a_s, c, a_p, k$ 
 $\mathbb{R}^5 \leftarrow \text{INC}$  is the array of increments for  $o, a_s, c, a_p, k$ 
if  $\text{rs} < 6$  then
  for each  $i$  in  $W$  do
    while  $\text{frs} < 15$  do
      if  $W(i) + \text{INC}(i) > \text{MAX}(i)$  then
         $W(i) = \text{MAX}(i)$ 
      else
         $W(i) \leftarrow W(i) + \text{INC}(i)$ 
      end if
       $\text{frs} \leftarrow \text{Sum}(W.\text{values})$ 
    end while
  end for
else if  $\text{rs} < 12$  then
  for each  $i$  in  $W$  do
    while  $\text{frs} < 20$  do
      if  $W(i) + \text{INC}(i) > \text{MAX}(i)$  then
         $W(i) = \text{MAX}(i)$ 
      else
         $W(i) \leftarrow W(i) + \text{INC}(i)$ 
      end if
       $\text{frs} \leftarrow \text{Sum}(W.\text{values})$ 
    end while
  end for
else if  $\text{rs} < 19$  then
  for each  $i$  in  $W$  do
    while  $\text{frs} > 2$  do
      if  $W(i) + \text{INC}(i) < \text{MIN}(i)$  then
         $W(i) = \text{MIN}(i)$ 
      else
         $W(i) \leftarrow W(i) - \text{INC}(i)$ 
      end if
       $\text{frs} \leftarrow \text{Sum}(W.\text{values})$ 
    end while
  end for
else if  $\text{rs} \geq 19$  then
  for each  $i$  in  $W$  do
    while  $\text{frs} > 9$  do
      if  $W(i) + \text{INC}(i) < \text{MIN}(i)$  then
         $W(i) = \text{MIN}(i)$ 
      else
         $W(i) \leftarrow W(i) - \text{INC}(i)$ 
      end if
       $\text{frs} \leftarrow \text{Sum}(W.\text{values})$ 
    end while
  end for
end if

```

Algorithm 2: The biased Robex algorithm used to make inappropriate recommendations and explanations.

C1. Workshop guide

Figure C.1: The following figures present the questions used in the workshops conducted in Chapter 6.

Each workshop included the following steps:

0. Participants read and fill in the consent form, and then the pre-questionnaire (paper format)
1. Participants are asked questions on the normal procedure in their AML-CFT profession (either the control procedures for regulators or conception procedure for model designers in banks)
2. Participants questions are asked questions about the use of AI in AML-CFT to understand their impressions on AI.
3. A scenario where AI is used in AML-CFT transaction monitoring systems is then introduced and participants are asked questions about this scenario.
4. Finally, conceptual design artifacts of different explanations and justifications are shown to participants. Participants are asked to discuss them.

Below are the questions asked to participants. The questions have been adapted slightly depending on whether they were asked to regulators or bank practitioners. Questions for regulators are shown in the **blue boxes**, those for participants from banks in the **red boxes**. The **purple boxes** indicate that there was no difference between the questions asked to regulators and bank practitioners for the phase in question.

Phase 0: Pre-Questionnaire (Regulators)

1. How many years of experience do you have in controlling AML/CFT systems? (Between 1 and 3, Between 4 and 10, More than 10 years)
2. Do you have any specific expertise in LCB-FT?
3. What is your level of familiarity with: artificial intelligence? The cloud? Big data? (Likert-type responses on a scale of 1 to 7)

Phase 0: Pre-Questionnaire (Banks)

1. How many years of experience do you have in AML/CFT systems in financial institutions? (Between 1 and 3, Between 4 and 10, More than 10 years)
2. Do you have any specific expertise in LCB-FT?
3. What is your level of familiarity with: artificial intelligence? The cloud? Big data? (Likert-type responses on a scale of 1 to 7)

Phase 1: Understanding the control processes (Regulators)

4. What are the different steps of a control? What are the criteria to evaluate AML/CFT processes?
5. What should banks justify/explain regarding the tools used in AML/CFT (the example of transaction monitoring could be used)?
6. What form do these justifications take?

Phase 1: Understanding the implementation of models in AML-CFT in banks

4. What are the different steps in implementing a financial security project?
5. What should banks justify/explain regarding the tools used in AML/CFT (the example of transaction monitoring could be used)?
6. What form do these justifications take?

Phase 2: Impressions on AI

1. What new technologies are emerging in banks' AML/CFT systems?
2. Can these situations be linked to artificial intelligence in your opinion: data collection, customer risk characterization, transaction monitoring system, alert review, monitoring tools. If so, what is the role of AI in these systems?
3. How promising do you think this technology is?
4. Do you think (and why) that using AI could be more or less risky for financial security than current systems (without AI)?
5. Do you think these systems could be more or less difficult to control/monitor?

AI debrief: at the end of this phase, if the participants are not very familiar with AI, the moderator will define AI (OECD and Wikipedia definitions of AI and Machine learning) and give a short presentation on different types of machine learning.

For phase 3, a "scenario" will be introduced. It describes a hypothetical situation involving an AI system in a bank. Its purpose is to provoke questions from the controllers and to bring out ideas. It also features a fictional character. The purpose of this character is to encourage the participants to immerse themselves in a situation and encourage them to speak freely and react to details.

There are two different scenarios involving AI in AML-CFT transaction monitoring systems:

- Case study 1: Automatic redirection and closing of alerts (Transaction Monitoring)
- Case study 2: Detection of new risk typologies (Transaction Monitoring)

See the scenarios in the rest of the registration files for more details.

Phase 3: The need for justifications

1. Do you think this use of AI is legitimate? useful?
2. What will Eric want to know to audit the system? What questions will Eric want to know about the algorithm?
3. Does Bank B have to justify the use and the potential added value of AI? If so, how? What would be the baseline?
4. Does Bank B need to justify changing or even eliminating any existing systems? If so, how?
5. Is it possible to set an overall system performance target in the AML-CFT environment? If so, how can it be quantified? If not, why not?

For phase 4, examples of justifications are shown to participants showing examples of explanations of AI systems/decisions.

Phase 4: Ideation on justifications

6. Are these justifications useful? Are they good ones? Are they necessary? Why?
7. What are the limits of these justifications? How can they be improved?

C2. Compliance assessment

Table C.2: Summary of the compliance assessment made in Chapter 6 to determine the points in the AML-CFT legislation with which AI opacity interferes. The assessment was made for the two AI use cases presented in Figure 6.3.1: "SR" refers to "Risk Scoring" (Scenario 1), and "NT" to "New typologies (scenario 2).

AML-CFT Theme	Legal reference	Is AI opacity a problem? For which model?	Why?
Customer knowledge and constant vigilance over business relationships	French Monetary Code (CMF) Articles L.561-4-1 to L. 561-14-2	No	The update of customer and beneficial owner databases is not made with AI in the use cases we are considering.
Risk classification	CMF Article L. 561-4-1	Yes for NT	Banks need to understand the new typologies of risk detected by the AI to update their risk classification.
Calibration / allocation of material and human resources	CMF Article R. 561-38	Yes for RS	Assessing the suitability of AI for prioritizing alerts
Constant vigilance	CMF Article L. 561-6	Yes for NT	Justifications might be needed on the training frequency.
Careful examination: Ability to detect inconsistencies/anomalies	CMF Article L. 561-6	Yes for NT	The relevance of a model can be justified with performance statistics, but understanding why an anomaly was not detected is important for both supervisors and banks.
Processing alerts in a timely manner	Sanction Decision BMW Finance 16/06/23	Yes for NT and SR	AI opacity can make reviews longer
Adaptation / completeness of the system in relation to the risk classification	CMF Article R. 561-12-1, Sanction Decision Axa Banque 15/02/23	Yes for NT	The alignment between human and machine on important parameters should be demonstrated
Enhanced vigilance: ability to analyze risky alerts	CMF Article L. 561-10-2	Yes for SR	We need to be able to understand the criteria that generate a risky alert.
SAR obligation: ability to produce high-quality SAR when relevant	CMF Article L. 561-15	Yes for SR and NT	We need to be able to understand the criteria that generate a risky alert.
Internal control: incident detection; Stability over time; mastering of the system (from external service provider); Safety net in case of failure	CMF Article R561-38-4, Order of November 3, 2014	Yes for SR and NT	Have to be able to anticipate the model's behavior to anticipate plausible incidents; Have to demonstrate AI behavior does not drift; Have to be able to demonstrate the control of your system.

Bibliography

- Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–18, New York, NY, USA, April 2018. Association for Computing Machinery. ISBN 978-1-4503-5620-6. DOI: 10.1145/3173574.3174156. URL <https://doi.org/10.1145/3173574.3174156>.
- Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y. Lim. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, New York, NY, USA, April 2020. Association for Computing Machinery. ISBN 978-1-4503-6708-0. URL <https://doi.org/10.1145/3313831.3376615>.
- Chadia Abras, Diane Maloney-Krichmar, Jenny Preece, and others. User-centered design. *Bainbridge, W. Encyclopedia of Human-Computer Interaction*. Thousand Oaks: Sage Publications, 37(4):445–456, 2004.
- Daron Acemoglu. Harms of AI, September 2021. URL <https://www.nber.org/papers/w29247>. Accessed 2023-11-27.
- Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, January 2015. DOI: 10.1126/science.aaa1465. URL <https://www.science.org/doi/10.1126/science.aaa1465>. Publisher: American Association for the Advancement of Science.
- A. Adadi and M. Berrada. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018. ISSN 2169-3536. DOI: 10.1109/ACCESS.2018.2870052. Conference Name: IEEE Access.
- Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity Checks for Saliency Maps, November 2020. URL <http://arxiv.org/abs/1810.03292>. arXiv:1810.03292 [cs, stat].
- Darius Afchar, Alessandro B. Melchiorre, Markus Schedl, Romain Hennequin, Elena V. Epure, and Manuel Moussallam. Explainability in music recommender systems. *AI Magazine*, 43(2):190–208, 2022. ISSN 2371-9621. DOI: 10.1002/aaai.12056. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/aaai.12056>. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/aaai.12056>.
- Sabbir Ahmad, Andy Bryant, Erica Kleinman, Zhaoqing Teng, Truong-Huy D. Nguyen, and Magy Seif El-Nasr. Modeling Individual and Team Behavior through Spatio-temporal Analysis. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '19, pages 601–612, New York, NY, USA, October 2019. Association for Computing Machinery. ISBN 978-1-4503-6688-5. DOI: 10.1145/3311350.3347188. URL <https://doi.org/10.1145/3311350.3347188>.
- Ahn, Yongsu, Yan, Muheng, Lin, Yu-Ru, Chung, Wen-Ting, and Hwa, Rebecca. Tribe or Not? Critical Inspection of Group Differences Using TribalGram. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, March 2022. DOI: 10.1145/3484509. URL <https://dl.acm.org/doi/full/10.1145/3484509>. Publisher: ACM PUB27 New York, NY.

- Zeynep Akata, Dan Balliet, Maarten de Rijke, Frank Dignum, Virginia Dignum, Gusztai Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, Hayley Hung, Catholijn Jonker, Christof Monz, Mark Neerincx, Frans Oliehoek, Henry Prakken, Stefan Schlobach, Linda van der Gaag, Frank van Harmelen, Herke van Hoof, Birna van Riemsdijk, Aimee van Wynsberghe, Rineke Verbrugge, Bart Verheij, Piek Vossen, and Max Welling. A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer*, 53(8):18–28, August 2020. ISSN 1558-0814. DOI: 10.1109/MC.2020.2996587. Conference Name: Computer.
- Ekin Akyürek, Tolga Bolukbasi, Frederick Liu, Binbin Xiong, Ian Tenney, Jacob Andreas, and Kelvin Guu. Towards Tracing Factual Knowledge in Language Models Back to the Training Data, October 2022. URL <http://arxiv.org/abs/2205.11482>. arXiv:2205.11482 [cs].
- Raghad Al-Shabandar, Gaye Lightbody, Fiona Browne, Jun Liu, Haiying Wang, and Huiru Zheng. The Application of Artificial Intelligence in Financial Compliance Management. In *Proceedings of the 2019 International Conference on Artificial Intelligence and Advanced Manufacturing*, AIAM 2019, pages 1–6, New York, NY, USA, October 2019. Association for Computing Machinery. ISBN 978-1-4503-7202-2. DOI: 10.1145/3358331.3358339. URL <https://dl.acm.org/doi/10.1145/3358331.3358339>.
- J Alammar. Ecco: An Open Source Library for the Explainability of Transformer Language Models. In Heng Ji, Jong C. Park, and Rui Xia, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 249–257, Online, August 2021. Association for Computational Linguistics. DOI: 10.18653/v1/2021.acl-demo.30. URL <https://aclanthology.org/2021.acl-demo.30>. Accessed 2023-11-02.
- Kars Alfrink, Ianus Keller, Neelke Doorn, and Gerd Kortuem. Contestable Camera Cars: A Speculative Design Exploration of Public AI That Is Open and Responsive to Dispute. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–16, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 978-1-4503-9421-5. DOI: 10.1145/3544548.3580984. URL <https://dl.acm.org/doi/10.1145/3544548.3580984>.
- R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 111–117, October 2005. DOI: 10.1109/INFVIS.2005.1532136. ISSN: 1522-404X.
- Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine*, 35(4):105–120, December 2014. ISSN 2371-9621. DOI: 10.1609/aimag.v35i4.2513. URL <https://ojs.aaai.org/index.php/aimagazine/article/view/2513>. Accessed 2021-04-16.
- Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Glasgow Scotland Uk, May 2019. ACM. ISBN 978-1-4503-5970-2. DOI: 10.1145/3290605.3300233. URL <https://dl.acm.org/doi/10.1145/3290605.3300233>.
- Jennifer Amsterlaw and Henry M. Wellman. Theories of Mind in Transition: A Microgenetic Study of the Development of False Belief Understanding. *Journal of Cognition and Development*, 7(2):139–172, 2006. ISSN 1532-7647. Place: US Publisher: Lawrence Erlbaum.
- Geoffrey R. Amthor. Multimedia in education: an introduction. *Int. Business Mag.*, pages 32–39, 1992. ISSN 0192-592X.
- Natalia Andrienko, Gennady Andrienko, Linara Adilova, Stefan Wrobel, and Theresa-Marie Rhyne. Visual Analytics for Human-Centered Machine Learning. *IEEE computer graphics and applications*, 42(1): 123–133, February 2022. ISSN 1558-1756. DOI: 10.1109/MCG.2021.3130314.

- Ariful Islam Anik and Andrea Bunt. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Yokohama Japan, May 2021. ACM. ISBN 978-1-4503-8096-6. DOI: 10.1145/3411764.3445736. URL <https://dl.acm.org/doi/10.1145/3411764.3445736>.
- Anthropic. Decomposing Language Models Into Understandable Components, October 2023. URL <http://www.anthropic.com/index/decomposing-language-models-into-understandable-components>. Accessed 2023-12-12.
- Hilary Arksey and Lisa O'Malley. Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology*, 8(1):19–32, February 2005. ISSN 1364-5579. DOI: 10.1080/1364557032000119616. URL <https://doi.org/10.1080/1364557032000119616>. Publisher: Routledge _eprint: <https://doi.org/10.1080/1364557032000119616>.
- Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. *arXiv:1909.03012 [cs, stat]*, September 2019. URL <http://arxiv.org/abs/1909.03012>. arXiv: 1909.03012.
- S. Atakishiyev, H. Babiker, N. Farruque, R. Goebel, M.-Y. Kima, M. H. Motallebi, J. Rabelo, T. Syed, and O. R. Zaïane. A multi-component framework for the analysis and design of explainable artificial intelligence, May 2020. URL <http://arxiv.org/abs/2005.01908>. arXiv:2005.01908 [cs].
- Akanksha Atrey, Kaleigh Clary, and David Jensen. Exploratory Not Explanatory: Counterfactual Analysis of Saliency Maps for Deep Reinforcement Learning, February 2020. URL <http://arxiv.org/abs/1912.05743>. arXiv:1912.05743 [cs].
- Autorité de Contrôle Prudentiel et de Résolution. Annual Report of the ACPR 2022. Technical report, ACPR, Bank of France, May 2023a. URL https://acpr.banque-france.fr/sites/default/files/medias/documents/20230524_rapport_annuel_colb_2022.pdf. Accessed 11/29/2023.
- Autorité de Contrôle Prudentiel et de Résolution. Thematic review on automated systems for monitoring AML/CFT transactions. Technical report, ACPR, Bank of France, April 2023b. URL <https://acpr.banque-france.fr/dispositifs-automatisees-de-surveillance-des-operations-en-matiere-de-lcb-ft>. Accessed 2023-08-26.
- Michel Avital and Dov Te'eni. From generative fit to generative capacity: exploring an emerging dimension of information systems design and task performance. *Information Systems Journal*, 19(4):345–367, 2009. ISSN 1365-2575. DOI: 10.1111/j.1365-2575.2007.00291.x. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2575.2007.00291.x>.
- S. Sandra Bae, Clement Zheng, Mary Etta West, Ellen Yi-Luen Do, Samuel Huron, and Danielle Albers Szafir. Making Data Tangible: A Cross-disciplinary Design Space for Data Physicalization. In *CHI Conference on Human Factors in Computing Systems*, pages 1–18, New Orleans LA USA, April 2022. ACM. ISBN 978-1-4503-9157-3. DOI: 10.1145/3491102.3501939. URL <https://dl.acm.org/doi/10.1145/3491102.3501939>.
- N. Bagheri and G. Jamieson. Considering subjective trust and monitoring behavior in assessing automation-induced “complacency”. 2004. URL <https://www.semanticscholar.org/paper/CONSIDERING-SUBJECTIVE-TRUST-AND-MONITORING-IN-Bagheri-Jamieson/4338960e130f8ddb57815b67f34c4a03264ab820>. Accessed 2024-01-09.
- N. R. Bailey and M. W. Scerbo. Automation-induced complacency for monitoring highly reliable systems: the role of task complexity, system experience, and operator trust. *Theoretical Issues in Ergonomics Science*, 8(4):321–348, July 2007. ISSN 1463-922X. DOI: 10.1080/14639220500535301. URL <https://doi.org/10.1080/14639220500535301>. _eprint: <https://doi.org/10.1080/14639220500535301> tex.ids=bailey_automation-induced_2007 publisher: Taylor & Francis.

- Lisanne Bainbridge. Ironies of automation. *Automatica*, 19(6):775–779, November 1983. ISSN 0005-1098. URL <https://www.sciencedirect.com/science/article/pii/0005109883900468>.
- Ramnath Balasubramanian, Ari Chester, and Nick Milinkovich. Rewriting the rules: Digital and AI-powered underwriting in life insurance. Consultancy Report, McKinsey & Company, July 2020. URL <https://www.mckinsey.com/industries/financial-services/our-insights/rewriting-the-rules-digital-and-ai-powered-underwriting-in-life-insurance>. Accessed 2023-01-31.
- Ramnath Balasubramanian, Ari Libarikian, and Doug McElhaney. Insurance 2030—The impact of AI on the future of insurance. Technical report, McKinsey & Company, March 2021. URL <https://www.mckinsey.com/industries/financial-services/our-insights/insurance-2030-the-impact-of-ai-on-the-future-of-insurance>. Accessed 2023-01-31.
- Agathe Balayn, Mireia Yurrita, Jie Yang, and Ujwal Gadiraju. Fairness Toolkits, A Checkbox Culture? On the Factors that Fragment Developer Practices in Handling Algorithmic Harms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, pages 482–495, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702310. DOI: 10.1145/3600211.3604674. URL <https://dl.acm.org/doi/10.1145/3600211.3604674>.
- Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–16, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8096-6. DOI: 10.1145/3411764.3445717. URL <https://doi.org/10.1145/3411764.3445717>.
- Natã M. Barbosa and Monchu Chen. Rehumanized Crowdsourcing: A Labeling Framework Addressing Bias and Ethics in Machine Learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–12, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 978-1-4503-5970-2. DOI: 10.1145/3290605.3300773. URL <https://doi.org/10.1145/3290605.3300773>.
- Pablo Barceló, Egor V Kostylev, Mikaël Monet, Jorge Pérez, Juan Reutter, and Juan-Pablo Silva. The logical expressiveness of graph neural networks. In *8th International Conference on Learning Representations (ICLR 2020)*, Virtual conference, Ethiopia, April 2020. URL <https://hal.science/hal-03356968>.
- Philip Barker. Designing Interactive Learning. In Ton de Jong and Luigi Sarti, editors, *Design and Production of Multimedia and Simulation-based Learning Material*, pages 1–30. Springer Netherlands, Dordrecht, 1994. ISBN 978-94-011-0942-0. URL https://doi.org/10.1007/978-94-011-0942-0_1.
- Solon Barocas, Andrew D. Selbst, and Manish Raghavan. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 80–89, New York, NY, USA, January 2020. Association for Computing Machinery. ISBN 978-1-4503-6936-7. DOI: 10.1145/3351095.3372830. URL <https://doi.org/10.1145/3351095.3372830>.
- Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, June 2020. ISSN 1566-2535. DOI: 10.1016/j.inffus.2019.12.012. URL <http://www.sciencedirect.com/science/article/pii/S1566253519308103>.
- Sarah Bayer, Henner Gimpel, and Moritz Markgraf. The role of domain expertise in trusting and following explainable AI decision support systems. *Journal of Decision Systems*, 0(0):1–29, 2021. ISSN 1246-0125. DOI: 10.1080/12460125.2021.1958505. URL <https://doi.org/10.1080/12460125.2021.1958505>. Publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/12460125.2021.1958505>.

- Valérie Beaudouin, Isabelle Bloch, David Bounie, Stéphan Cléménçon, Florence d’Alché Buc, James Eagan, Winston Maxwell, Pavlo Mozharovskiy, and Jayneel Parekh. Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach, March 2020. URL <http://arxiv.org/abs/2003.07703>. arXiv:2003.07703 [cs].
- Sander Beckers. Causal Explanations and XAI, February 2022. URL <http://arxiv.org/abs/2201.13169>. arXiv:2201.13169 [cs].
- Esube Bekele, Wallace E. Lawson, Zachary Horne, and Sangeet Khemlani. Implementing a Robust Explanatory Bias in a Person Re-identification Network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2246–22467, June 2018. DOI: 10.1109/CVPRW.2018.00291. ISSN: 2160-7516.
- R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. Natesan Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4:1–4:15, July 2019. ISSN 0018-8646. DOI: 10.1147/JRD.2019.2942287. Conference Name: IBM Journal of Research and Development.
- Luigi Bellomarini, Eleonora Laurenza, and Emanuel Sallinger. Rule-based Anti-Money Laundering in Financial Intelligence Units: Experience and Vision. In *Proceedings of the 14th International Rule Challenge, 4th Doctoral Consortium, and 6th Industry Track @ RuleML+RR 2020*, page 12, Oslo, Norway, July 2020. CEUR Workshop Proceedings.
- Mariette Bengtsson. How to plan and perform a qualitative study using content analysis. *NursingPlus Open*, 2:8–14, January 2016. ISSN 2352-9008. DOI: 10.1016/j.npls.2016.01.001. URL <https://www.sciencedirect.com/science/article/pii/S2352900816000029>.
- Astrid Bertrand, Winston Maxwell, and Xavier Vamparys. Do AI-based anti-money laundering (AML) systems violate European fundamental rights? *International Data Privacy Law*, 11(3):276–293, August 2021. ISSN 2044-3994. DOI: 10.1093/idpl/ipab010. URL <https://doi.org/10.1093/idpl/ipab010>.
- Astrid Bertrand, Rafik Belloum, James R. Eagan, and Winston Maxwell. How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’22, pages 78–91, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 978-1-4503-9247-1. DOI: 10.1145/3514094.3534164. URL <https://doi.org/10.1145/3514094.3534164>.
- Cornelia Betsch, Niels Haase, Frank Renkewitz, and Philipp Schmid. The narrative bias revisited: What drives the biasing influence of narrative information on risk perceptions? *Judgment and Decision Making*, 10(3):241–264, May 2015. ISSN 1930-2975. DOI: 10.1017/S1930297500004654. URL <https://www.cambridge.org/core/journals/judgment-and-decision-making/article/narrative-bias-revisited-what-drives-the-biasing-influence-of-narrative-information-on-risk-perceptions/52E778EFD11CA174B5573A4AFE3664E1>. Publisher: Cambridge University Press.
- Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* ’20, pages 648–657, New York, NY, USA, January 2020. Association for Computing Machinery. ISBN 978-1-4503-6936-7. DOI: 10.1145/3351095.3375624. URL <https://doi.org/10.1145/3351095.3375624>.
- Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunnara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 401–413, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8473-5. URL <https://doi.org/10.1145/3461702.3462571>.

- Adrien Bibal, Michael Lognoul, Alexandre de Streel, and Benoît Frénay. Legal requirements on explainability in machine learning. *Artificial Intelligence and Law*, 29(2):149–169, June 2021. ISSN 1572-8382. DOI: 10.1007/s10506-020-09270-4. URL <https://doi.org/10.1007/s10506-020-09270-4>.
- Charles E. Billings. *Human-Centered Aviation Automation: Principles and Guidelines*. NASA Technical Memorandum, February 1996.
- Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. Impossibility Theorems for Feature Attribution, April 2023. URL <http://arxiv.org/abs/2212.11870>. arXiv:2212.11870 [cs].
- Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. ‘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*, CHI ’18, pages 1–14, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 978-1-4503-5620-6. DOI: 10.1145/3173574.3173951. URL <https://doi.org/10.1145/3173574.3173951>.
- Julia Black. Decentring Regulation: Understanding the Role of Regulation and Self-Regulation in a ‘Post-Regulatory’ World. *Current Legal Problems*, 54(1):103–146, January 2001. ISSN 0070-1998. DOI: 10.1093/clp/54.1.103. URL <https://doi.org/10.1093/clp/54.1.103>.
- Douglas Blakey. AI in anti money laundering, December 2022. URL <https://www.retailbankerinternational.com/comment/ai-money-laundering/>. Accessed 2023-08-26.
- Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation, Financial Crimes Enforcement Network, National Credit Union Administration, and Office of the Comptroller of the Currency. Joint Statement on Innovative Efforts to Combat Money Laundering and Terrorist Financing. Technical report, Federal Reserve Board, December 2018. URL <https://www.federalreserve.gov/newsevents/pressreleases/files/bcreg20181203a1.pdf>. Accessed 11/19/2023.
- Margaret A. Boden. *Artificial Intelligence*. Elsevier, June 1996. ISBN 978-0-08-052759-8. Google-Books-ID: _ixmRlL0jclC.
- Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. Contextualization and Exploration of Local Feature Importance Explanations to Improve Understanding and Satisfaction of Non-Expert Users. In *27th International Conference on Intelligent User Interfaces*, pages 807–819, Helsinki Finland, March 2022. ACM. ISBN 978-1-4503-9144-3. DOI: 10.1145/3490099.3511139. URL <https://dl.acm.org/doi/10.1145/3490099.3511139>.
- John Braithwaite and Toni Makkai. Trust and compliance. *Policing and Society*, 4(1):1–12, May 1994. ISSN 1043-9463. DOI: 10.1080/10439463.1994.9964679. URL <https://doi.org/10.1080/10439463.1994.9964679>. Publisher: Routledge _eprint: <https://doi.org/10.1080/10439463.1994.9964679>.
- Laura Brandimarte, Alessandro Acquisti, and George Loewenstein. Misplaced Confidences: Privacy and the Control Paradox. *Social Psychological and Personality Science*, 4(3):340–347, May 2013. ISSN 1948-5506. DOI: 10.1177/1948550612455931. URL <https://doi.org/10.1177/1948550612455931>. Publisher: SAGE Publications Inc.
- Dawn Branley-Bell, Rebecca Whitworth, and Lynne Coventry. User Trust and Understanding of Explainable AI: Exploring Algorithm Visualisations and User Biases. In *Human-Computer Interaction. Human Values and Quality of Life: Thematic Area, HCI 2020, Held as Part of the 22nd International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part III*, pages 382–399, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-49064-5. URL https://doi.org/10.1007/978-3-030-49065-2_27.
- David A Broniatowski. Psychological foundations of explainability and interpretability in artificial intelligence. Technical Report NIST IR 8367, National Institute of Standards and Technology (U.S.), Gaithersburg, MD, April 2021. URL <https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8367.pdf>.

- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, July 2020. URL <http://arxiv.org/abs/2005.14165>. arXiv:2005.14165 [cs].
- Joanna J. Bryson and Andreas Theodorou. How Society Can Maintain Human-Centric Artificial Intelligence. In Marja Toivonen and Eveliina Saari, editors, *Human-Centered Digitalization and Services*, Transnational Systems Sciences, pages 305–323. Springer Nature, Singapore, 2019. ISBN 9789811377259. URL https://doi.org/10.1007/978-981-13-7725-9_16.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of Artificial General Intelligence: Early experiments with GPT-4, April 2023. URL <http://arxiv.org/abs/2303.12712>. arXiv:2303.12712 [cs].
- B.G. Buchanan and E.H. Shortliffe. *Rule-based expert systems: The mycin experiments of the stanford heuristic programming project: B.G. Buchanan and E.H. Shortliffe*. Addison-Wesley, Reading, MA, July 1985.
- Mario Bunge. *Philosophy of Science: From Explanation to Justification*. Transaction Publishers, 1998. ISBN 978-1-4128-3083-6. Google-Books-ID: ofcy8wZeLCoC.
- Nadia Burkart and Marco F. Huber. A Survey on the Explainability of Supervised Machine Learning. *Journal of Artificial Intelligence Research*, 70:245–317, January 2021. ISSN 1076-9757. DOI: 10.1613/jair.1.12228. URL <http://arxiv.org/abs/2011.07876>.
- Jenna Burrell. How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1):2053951715622512, June 2016. ISSN 2053-9517. DOI: 10.1177/2053951715622512. URL <https://doi.org/10.1177/2053951715622512>. Publisher: SAGE Publications Ltd.
- A. Bussone, S. Stumpf, and D. O’Sullivan. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In *2015 International Conference on Healthcare Informatics*, pages 160–169, October 2015. DOI: 10.1109/ICHI.2015.26.
- Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 454–464, Cagliari Italy, March 2020. ACM. ISBN 978-1-4503-7118-6. DOI: 10.1145/3377325.3377498. URL <https://dl.acm.org/doi/10.1145/3377325.3377498>.
- Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):188:1–188:21, 2021. DOI: 10.1145/3449287. URL <https://doi.org/10.1145/3449287>.
- Christoph Bösch, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher. Tales from the Dark Side: Privacy Dark Strategies and Privacy Dark Patterns. *Proceedings on Privacy Enhancing Technologies*, 2016. ISSN 2299-0984. URL <https://petsymposium.org/popets/2016/popets-2016-0038.php>.
- Susanne Bødker. When second wave HCI meets third wave challenges. In *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles*, pages 1–8, Oslo Norway, October 2006. ACM. ISBN 978-1-59593-325-6. DOI: 10.1145/1182475.1182476. URL <https://dl.acm.org/doi/10.1145/1182475.1182476>.
- Ángel Alexander Cabrera, Marco Tulio Ribeiro, Bongshin Lee, Rob DeLine, Adam Perer, and Steven M Drucker. What Did My AI Learn? How Data Scientists Make Sense of Model Behavior. *ACM Transactions on Computer-Human Interaction*, 2022. Publisher: ACM New York, NY.

- Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI '19*, pages 258–262, New York, NY, USA, March 2019. Association for Computing Machinery. ISBN 978-1-4503-6272-6. DOI: 10.1145/3301275.3302289. URL <https://dl.acm.org/doi/10.1145/3301275.3302289>.
- Michael L. Callaham, Robert L. Wears, Ellen J. Weber, Christopher Barton, and Gary Young. Positive-Outcome Bias and Other Limitations in the Outcome of Research Abstracts Submitted to a Scientific Meeting. *JAMA*, 280(3):254–257, July 1998. ISSN 0098-7484. DOI: 10.1001/jama.280.3.254. URL <https://doi.org/10.1001/jama.280.3.254>.
- Ana Isabel Canhoto. Leveraging machine learning in the global fight against money laundering and terrorism financing: An affordances perspective. *Journal of Business Research*, 131:441–452, October 2020. ISSN 0148-2963. DOI: 10.1016/j.jbusres.2020.10.012. URL <http://www.sciencedirect.com/science/article/pii/S0148296320306640>.
- John M. Carroll. Chapter 17 - Scenario-Based Design. In Marting G. Helander, Thomas K. Landauer, and Prasad V. Prabhu, editors, *Handbook of Human-Computer Interaction (Second Edition)*, pages 383–406. North-Holland, Amsterdam, January 1997. ISBN 978-0-444-81862-1. DOI: 10.1016/B978-044481862-1.50083-2. URL <https://www.sciencedirect.com/science/article/pii/B9780444818621500832>.
- Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, and Chris Olah. Activation Atlas. *Distill*, 4(3):e15, March 2019. ISSN 2476-0757. DOI: 10.23915/distill.00015. URL <https://distill.pub/2019/activation-atlas>.
- Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '15*, pages 1721–1730, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 978-1-4503-3664-2. DOI: 10.1145/2783258.2788613. URL <https://dl.acm.org/doi/10.1145/2783258.2788613>.
- Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, 8(8):832, August 2019. DOI: 10.3390/electronics8080832. URL <https://www.mdpi.com/2079-9292/8/8/832>. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.
- A. Cawsey. User modelling in interactive explanations. *User Modeling and User-Adapted Interaction*, 3(3): 221–247, 1993. ISSN 1573-1391. DOI: 10.1007/BF01257890. Springer.
- Raja Chatila, Virginia Dignum, Michael Fisher, Fosca Giannotti, Katharina Morik, Stuart Russell, and Karen Yeung. Trustworthy AI. In Bertrand Braunschweig and Malik Ghallab, editors, *Reflections on Artificial Intelligence for Humanity*, Lecture Notes in Computer Science, pages 13–39. Springer International Publishing, Cham, 2021. URL https://doi.org/10.1007/978-3-030-69128-8_2.
- Aditya Chattopadhyay, Piyushi Manupriya, Anirban Sarkar, and Vineeth N. Balasubramanian. Neural Network Attributions: A Causal Perspective, July 2019. URL <http://arxiv.org/abs/1902.02302>. arXiv:1902.02302 [cs, stat].
- Larissa Chazette and Kurt Schneider. Explainability as a non-functional requirement: challenges and recommendations. *Requirements Engineering*, 25(4):493–514, December 2020. ISSN 1432-010X. DOI: 10.1007/s00766-020-00333-1. URL <https://doi.org/10.1007/s00766-020-00333-1>.
- Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This Looks Like That: Deep Learning for Interpretable Image Recognition. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL <https://proceedings.neurips.cc/paper/2019/hash/adf7ee2dcf142b0e11888e72b43fcb75-Abstract.html>.

- Valerie Chen, Q. Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):370:1–370:32, October 2023. DOI: 10.1145/3610219. URL <https://dl.acm.org/doi/10.1145/3610219>.
- Zhiyuan Chen, Le Dinh Van Khoa, Ee Na Teoh, Amril Nazir, Ettikan Kandasamy Karuppiah, and Kim Sim Lam. Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: a review. *Knowledge and Information Systems*, 57(2):245–285, November 2018. ISSN 0219-3116. DOI: 10.1007/s10115-017-1144-z. URL <https://doi.org/10.1007/s10115-017-1144-z>.
- Furui Cheng, Yao Ming, and Huamin Qu. DECE: Decision Explorer with Counterfactual Explanations for Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1438–1447, February 2021. ISSN 1941-0506. DOI: 10.1109/TVCG.2020.3030342. Conference Name: IEEE Transactions on Visualization and Computer Graphics.
- Furui Cheng, Dongyu Liu, Fan Du, Yanna Lin, Alexandra Zyttek, Haomin Li, Huamin Qu, and Kalyan Veeramachaneni. VBridge: Connecting the Dots Between Features and Data to Explain Healthcare Models. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):378–388, January 2022. ISSN 1941-0506. DOI: 10.1109/TVCG.2021.3114836. Conference Name: IEEE Transactions on Visualization and Computer Graphics.
- Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O’Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 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*, CHI ’19, pages 1–12, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 978-1-4503-5970-2. DOI: 10.1145/3290605.3300789. URL <https://doi.org/10.1145/3290605.3300789>.
- Micheline T. H. Chi, Nicholas De Leeuw, Mei-Hung Chiu, and Christian Lavancher. Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3):439–477, July 1994. ISSN 0364-0213. DOI: 10.1016/0364-0213(94)90016-7. URL <https://www.sciencedirect.com/science/article/pii/0364021394900167>.
- Michael Chromik and Andreas Butz. Human-XAI Interaction: A Review and Design Principles for Explanation User Interfaces. In Carmelo Ardito, Rosa Lanzilotti, Alessio Malizia, Helen Petrie, Antonio Piccinno, Giuseppe Desolda, and Kori Inkpen, editors, *Human-Computer Interaction – INTERACT 2021*, Lecture Notes in Computer Science, pages 619–640, Cham, 2021. Springer International Publishing. ISBN 978-3-030-85616-8.
- Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In *26th International Conference on Intelligent User Interfaces*, IUI ’21, pages 307–317, New York, NY, USA, April 2021. Association for Computing Machinery. ISBN 978-1-4503-8017-1. DOI: 10.1145/3397481.3450644. URL <https://doi.org/10.1145/3397481.3450644>.
- Douglas Cirqueira, Dietmar Nedbal, Markus Helfert, and Marija Bezbradica. Scenario-Based Requirements Elicitation for User-Centric Explainable AI. In Andreas Holzinger, Peter Kieseberg, A Min Tjoa, and Edgar Weippl, editors, *Machine Learning and Knowledge Extraction*, Lecture Notes in Computer Science, pages 321–341, Cham, 2020. Springer International Publishing. ISBN 978-3-030-57321-8.
- Ludovik Coba, Laurens Rook, Markus Zanker, and Panagiotis Symeonidis. Decision making strategies differ in the presence of collaborative explanations: two conjoint studies. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI ’19, pages 291–302, New York, NY, USA, March 2019. Association for Computing Machinery. ISBN 978-1-4503-6272-6. DOI: 10.1145/3301275.3302304. URL <https://doi.org/10.1145/3301275.3302304>.
- Dennis Collaris and Jarke J. van Wijk. ExplainExplore: Visual Exploration of Machine Learning Explanations. In *2020 IEEE Pacific Visualization Symposium (PacificVis)*, pages 26–35, June 2020. DOI: 10.1109/PacificVis48177.2020.7090. ISSN: 2165-8773.

- Jason A. Colquitt and Jessica B. Rodell. Measuring justice and fairness. In *The Oxford handbook of justice in the workplace*, Oxford library of psychology, pages 187–202. Oxford University Press, New York, NY, US, 2015. ISBN 978-0-19-998141-0. DOI: 10.1093/oxfordhb/9780199981410.013.8.
- Roberto Confalonieri, Ludovik Coba, Benedikt Wagner, and Tarek R. Besold. A historical perspective of explainable Artificial Intelligence. *WIREs Data Mining and Knowledge Discovery*, 11(1):e1391, 2021. ISSN 1942-4795. DOI: 10.1002/widm.1391. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1391>. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1391>.
- Conseil d’Orientation pour la lutte contre le blanchiment et le financement du terrorisme. Annual Report 2022. Technical report, COLB, May 2023. URL https://acpr.banque-france.fr/sites/default/files/medias/documents/20230524_rapport_annuel_colb_2022.pdf.
- Sven Coppers, Jan Van den Bergh, Kris Luyten, Karin Coninx, Iulianna van der Lek-Ciudin, Tom Vanallemeersch, and Vincent Vandeghinste. Intellingo: An Intelligible Translation Environment. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI ’18, pages 1–13, New York, NY, USA, April 2018. Association for Computing Machinery. ISBN 978-1-4503-5620-6. DOI: 10.1145/3173574.3174098. URL <https://doi.org/10.1145/3173574.3174098>.
- Juliet Corbin and Anselm Strauss. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. SAGE Publications, Inc. 2455 Teller Road Thousand Oaks, California 91320, November 2014. ISBN 978-1-4833-1568-3. Google-Books-ID: hZ6kBQAAQBAJ.
- Council of Europe. History of Artificial Intelligence - Artificial Intelligence - www.coe.int, 2023. URL <https://www.coe.int/en/web/artificial-intelligence/history-of-ai>.
- David Daniel Cox and Thomas Dean. Neural Networks and Neuroscience-Inspired Computer Vision. *Current Biology*, 24(18):R921–R929, September 2014. ISSN 0960-9822. DOI: 10.1016/j.cub.2014.08.026. URL <https://www.sciencedirect.com/science/article/pii/S0960982214010392>.
- Mark Craven and Jude Shavlik. Extracting Tree-Structured Representations of Trained Networks. In *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1995. URL <https://proceedings.neurips.cc/paper/1995/hash/45f31d16b1058d586fc3be7207b58053-Abstract.html>.
- John W. Creswell. *Qualitative Inquiry and Research Design: Choosing Among Five Approaches*. SAGE Publications, March 2012. ISBN 978-1-4129-9530-6. Google-Books-ID: OJYEbDtkxq8C.
- Michael Cui. The state of AI in 2023: Generative AI’s breakout year, April 2023. URL <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year>.
- Kimberly Culley and Poornima Madhavan. Trust in automation and automation designers: Implications for HCI and HMI. *Computers in Human Behavior*, 29(6):2208–2210, November 2013. ISSN 0747-5632. DOI: 10.1016/j.chb.2013.04.032. URL <http://www.sciencedirect.com/science/article/pii/S0747563213001441>.
- M. L. Cummings. Automation Bias in Intelligent Time Critical Decision Support Systems. In *AIAA 3rd Intelligent Systems Conference*, pages 2004–6313. AIAA, 2004.
- Brigham Daniels, Mark Buntaine, and Tanner Bangerter. Testing Transparency. *Northwestern University Law Review*, 114:1263, 2019. URL <https://heinonline.org/HOL/Page?handle=hein.journals/illlr114&id=1293&div=&collection=>.
- David Danks. The Value of Trustworthy AI. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 521–522, Honolulu HI USA, January 2019. ACM. ISBN 978-1-4503-6324-2. DOI: 10.1145/3306618.3314228. URL <https://dl.acm.org/doi/10.1145/3306618.3314228>.

- Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. Wearable Reasoner: Towards Enhanced Human Rationality Through A Wearable Device With An Explainable AI Assistant. In *Proceedings of the Augmented Humans International Conference, AHs '20*, pages 1–12, New York, NY, USA, March 2020. Association for Computing Machinery. ISBN 978-1-4503-7603-7. DOI: 10.1145/3384657.3384799. URL <https://doi.org/10.1145/3384657.3384799>.
- Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23*, pages 1–13, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 978-1-4503-9421-5. DOI: 10.1145/3544548.3580672. URL <https://dl.acm.org/doi/10.1145/3544548.3580672>.
- Arun Das and Paul Rad. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey, June 2020. URL <http://arxiv.org/abs/2006.11371>. arXiv:2006.11371 [cs].
- Richard Dazeley, Peter Vamplew, Cameron Foale, Charlotte Young, Sunil Aryal, and Francisco Cruz. Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artificial Intelligence*, 299:103525, October 2021. ISSN 0004-3702. DOI: 10.1016/j.artint.2021.103525. URL <https://www.sciencedirect.com/science/article/pii/S000437022100076X>.
- John Dewey. Democracy in Education. *THE ELEMENTARY SCHOOL TEACHER*, page 12, 1903.
- Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126, 2015. ISSN 1939-2222, 0096-3445. DOI: 10.1037/xge0000033. URL <http://doi.apa.org/getdoi.cfm?doi=10.1037/xge0000033>.
- Alan Dix and Geoffrey Ellis. Starting simple: adding value to static visualisation through simple interaction. In *Proceedings of the working conference on Advanced visual interfaces, AVI '98*, pages 124–134, New York, NY, USA, 1998. Association for Computing Machinery. ISBN 978-1-4503-7435-4. DOI: 10.1145/948496.948514. URL <https://doi.org/10.1145/948496.948514>.
- David H. Dodd and Jeffrey M. Bradshaw. Leading Questions and Memory: Pragmatic Constraints. *Journal of Verbal Learning and Verbal Behavior*, 19(6):695–704, December 1980. ERIC Number: EJ236855.
- Jonathan Dodge, Andrew A. Anderson, Matthew Olson, Rupika Dikkala, and Margaret Burnett. How Do People Rank Multiple Mutant Agents? In *27th International Conference on Intelligent User Interfaces, IUI '22*, pages 191–211, New York, NY, USA, March 2022. Association for Computing Machinery. ISBN 978-1-4503-9144-3. DOI: 10.1145/3490099.3511115. URL <https://doi.org/10.1145/3490099.3511115>.
- Vicente Dominguez, Pablo Messina, Ivania Donoso-Guzmán, and Denis Parra. The effect of explanations and algorithmic accuracy on visual recommender systems of artistic images. In *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI '19*, pages 408–416, New York, NY, USA, March 2019. Association for Computing Machinery. ISBN 978-1-4503-6272-6. DOI: 10.1145/3301275.3302274. URL <https://doi.org/10.1145/3301275.3302274>.
- Derek Doran, Sarah Schulz, and Tarek R. Besold. What Does Explainable AI Really Mean? A New Conceptualization of Perspectives, October 2017. URL <http://arxiv.org/abs/1710.00794>. arXiv:1710.00794 [cs].
- Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning, March 2017. URL <http://arxiv.org/abs/1702.08608>. arXiv:1702.08608 [cs, stat].
- Finale Doshi-Velez and Mason A. Kortz. Accountability of AI Under the Law: The Role of Explanation. *Berkman Klein Center for Internet & Society working paper*, Berkman Klein Center Working Group on Explanation and the Law:17, 2017. URL <https://dash.harvard.edu/handle/1/34372584>. Accepted: 2017-11-21T16:33:48Z Publisher: Berkman Klein Center for Internet & Society.

- Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0210–0215, May 2018. DOI: 10.23919/MIPRO.2018.8400040.
- Laurent Dupont, Olivier Fliche, and Su Yang. Governance of Artificial Intelligence in Finance. Discussion document, ACPR, June 2020.
- Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark Riedl. Automated Rationale Generation: A Technique for Explainable AI and its Effects on Human Perceptions. *arXiv:1901.03729 [cs]*, January 2019. URL <http://arxiv.org/abs/1901.03729>. arXiv: 1901.03729.
- Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. Expanding Explainability: Towards Social Transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–19, Yokohama Japan, May 2021. ACM. ISBN 978-1-4503-8096-6. DOI: 10.1145/3411764.3445188. URL <https://dl.acm.org/doi/10.1145/3411764.3445188>.
- Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. The Impact of Placebic Explanations on Trust in Intelligent Systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, pages 1–6, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 978-1-4503-5971-9. DOI: 10.1145/3290607.3312787. URL <https://doi.org/10.1145/3290607.3312787>.
- Malin Eiband, Daniel Buschek, and Heinrich Hussmann. How to Support Users in Understanding Intelligent Systems? Structuring the Discussion. In *26th International Conference on Intelligent User Interfaces*, IUI '21, pages 120–132, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8017-1. DOI: 10.1145/3397481.3450694. URL <https://doi.org/10.1145/3397481.3450694>.
- Satu Elo and Helvi Kyngäs. The qualitative content analysis process. *Journal of Advanced Nursing*, 62(1):107–115, 2008. ISSN 1365-2648. DOI: 10.1111/j.1365-2648.2007.04569.x. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2648.2007.04569.x>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2648.2007.04569.x>.
- Emmanuel Schizas, Grigory McKain, Bryan Zhang, Altantsetseg Ganbold, Pankajesh Kumar, Hatim Hussain, Kieran James Garvey, Eva Huang, Alexander Huang, Shaoxin Wang, and Nikos Yerolemou. The Global RegTech Industry Benchmark Report. Technical report, Cambridge Centre of Alternative Finance, 2019. URL <https://www.jbs.cam.ac.uk/wp-content/uploads/2020/08/2019-12-ccaf-global-regtech-benchmarking-report.pdf>.
- European Banking Authority. Guidelines on risk based supervision. Technical report, EBA, November 2016. URL <https://www.eba.europa.eu/regulation-and-policy/anti-money-laundering-and-e-money/guidelines-on-risk-based-supervision>.
- European Commission. Building Trust in Human-Centric Artificial Intelligence. Technical Report COM(2019) 168 final, August 2019. URL <https://digital-strategy.ec.europa.eu/en/library/communication-building-trust-human-centric-artificial-intelligence>.
- European Commission. Proposal for a Regulation of the European Parliament and of the Council laying down Harmonised Rules on Artificial Intelligence and amending certain Union Legislative Acts, April 2021. URL <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206>.
- European Commission. A European approach to artificial intelligence | Shaping Europe's digital future, October 2023. URL <https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence>.
- European Parliament and Council. Regulation (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) (Text with EEA relevance), April 2016. URL <http://data.europa.eu/eli/reg/2016/679/oj/eng>. Legislative Body: EP, CONSIL.

- European Parliament and Council. Regulation (EU) 2019/1150 of the European Parliament and of the Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services (Text with EEA relevance), June 2019. URL <http://data.europa.eu/eli/reg/2019/1150/oj/eng>. Legislative Body: EP, CONSIL.
- European Parliament and Council. Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online (Text with EEA relevance), April 2021. URL <http://data.europa.eu/eli/reg/2021/784/oj/eng>. Legislative Body: EP, CONSIL.
- European Parliament and Council. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Text with EEA relevance), October 2022. URL <http://data.europa.eu/eli/reg/2022/2065/oj/eng>. Legislative Body: EP, CONSIL.
- Chris Evans and Nicola J. Gibbons. The interactivity effect in multimedia learning. *Computers & Education*, 49(4):1147–1160, December 2007. ISSN 0360-1315. DOI: 10.1016/j.compedu.2006.01.008. URL <https://www.sciencedirect.com/science/article/pii/S0360131506000285>.
- Gregory Falco, Ben Shneiderman, Julia Badger, Ryan Carrier, Anton Dahbura, David Danks, Martin Eling, Alwyn Goodloe, Jerry Gupta, Christopher Hart, Marina Jirotko, Henric Johnson, Cara LaPointe, Ashley J. Llorens, Alan K. Mackworth, Carsten Maple, Sigurður Emil Pálsson, Frank Pasquale, Alan Winfield, and Zee Kin Yeong. Governing AI safety through independent audits. *Nature Machine Intelligence*, 3(7):566–571, July 2021. ISSN 2522-5839. DOI: 10.1038/s42256-021-00370-7. URL <https://www.nature.com/articles/s42256-021-00370-7>. Number: 7 Publisher: Nature Publishing Group.
- Andrea Falcon. Aristotle on Causality. January 2006. URL <https://plato.stanford.edu/ENTRIES/aristotle-causality/>. Last Modified: 2023-03-07.
- Peter Farley. Spotlight On Compliance Costs As Banks Get Down To Business With AI. *International Banker*, July 2017. URL <https://www.bankingexchange.com/bsa-aml/item/8202-cost-of-compliance-expected-to-hit-181bn>. Accessed 6/15/2020.
- Massimo Felici, Theofrastos Koulouris, and Siani Pearson. Accountability for Data Governance in Cloud Ecosystems. In *2013 IEEE 5th International Conference on Cloud Computing Technology and Science*, volume 2, pages 327–332, Bristol, UK, December 2013. IEEE. DOI: 10.1109/CloudCom.2013.157. URL <https://ieeexplore.ieee.org/abstract/document/6735445>. Accessed 12/12/2023.
- Shi Feng and Jordan Boyd-Graber. What can AI do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI '19*, pages 229–239, New York, NY, USA, March 2019. Association for Computing Machinery. ISBN 978-1-4503-6272-6. DOI: 10.1145/3301275.3302265. URL <https://doi.org/10.1145/3301275.3302265>.
- Andrea Ferrario and Michele Loi. How Explainability Contributes to Trust in AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, pages 1457–1466, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 978-1-4503-9352-2. DOI: 10.1145/3531146.3533202. URL <https://dl.acm.org/doi/10.1145/3531146.3533202>.
- Andrea Ferrario, Michele Loi, and Eleonora Viganò. In AI We Trust Incrementally: a Multi-layer Model of Trust to Analyze Human-Artificial Intelligence Interactions. *Philosophy & Technology*, 33(3):523–539, September 2020. ISSN 2210-5441. DOI: 10.1007/s13347-019-00378-3. URL <https://doi.org/10.1007/s13347-019-00378-3>.
- Juliana J. Ferreira and Mateus S. Monteiro. What Are People Doing About XAI User Experience? A Survey on AI Explainability Research and Practice. In Aaron Marcus and Elizabeth Rosenzweig, editors, *Design, User Experience, and Usability. Design for Contemporary Interactive Environments*, Lecture Notes in Computer Science, pages 56–73, Cham, 2020. Springer International Publishing. ISBN 978-3-030-49760-6.

- Financial Action Task Force. Guidance on the risk-based approach to combating money-laundering and terrorist financing. Technical report, FATF, June 2007. URL <https://www.fatf-gafi.org/en/publications/Fatfrecommendations/Fatfguidanceontherisk-basedapproachtocombatingmoneylaunderingandterroristfinancing-highlevelprinciplesandprocedures.html>. Accessed 12/2/2023.
- Financial Action Task Force. Risk-Based Approach for the Banking Sector. Technical report, FATF, October 2014. URL <https://www.fatf-gafi.org/en/publications/Fatfrecommendations/Risk-based-approach-banking-sector.html>. Accessed 12/02/2023.
- Financial Conduct Authority. Machine learning in UK financial services. Technical report, FCA, Bank of England, October 2019. URL <https://www.bankofengland.co.uk/-/media/boe/files/report/2019/machine-learning-in-uk-financial-services.pdf>.
- Financial Conduct Authority. Artificial Intelligence and Machine Learning. Technical Report DP-5-22, FCA, Bank of England, October 2022. URL <https://www.bankofengland.co.uk/-/media/boe/files/prudential-regulation/publication/2022/dp5-22--artificial-intelligence-and-machine-learning.pdf>.
- Financial Stability Board. Artificial intelligence and machine learning in financial services. Technical report, FSB, January 2017. URL <https://www.fsb.org/wp-content/uploads/P011117.pdf>.
- James D. Foley, Foley Dan Van, Andries Van Dam, Steven K. Feiner, and John F. Hughes. *Computer Graphics: Principles and Practice*. Addison-Wesley Professional, 1996. ISBN 978-0-201-84840-3.
- Sebastian Fritz-Morgenthal, Bernhard Hein, and Jochen Papenbrock. Financial Risk Management and Explainable, Trustworthy, Responsible AI. *Frontiers in Artificial Intelligence*, 5:14, 2022. ISSN 2624-8212. URL <https://www.frontiersin.org/articles/10.3389/frai.2022.779799>.
- Johannes Fürnkranz, Tomáš Kliegr, and Heiko Paulheim. On cognitive preferences and the plausibility of rule-based models. *Machine Learning*, 109(4):853–898, April 2020. ISSN 1573-0565. DOI: 10.1007/s10994-019-05856-5. URL <https://doi.org/10.1007/s10994-019-05856-5>.
- Krzysztof Z. Gajos and Lena Mamykina. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In *27th International Conference on Intelligent User Interfaces, IUI '22*, pages 794–806, New York, NY, USA, March 2022. Association for Computing Machinery. ISBN 978-1-4503-9144-3. DOI: 10.1145/3490099.3511138. URL <https://doi.org/10.1145/3490099.3511138>.
- Bill Gaver and Heather Martin. Alternatives: exploring information appliances through conceptual design proposals. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems, CHI '00*, pages 209–216, New York, NY, USA, 2000. Association for Computing Machinery. ISBN 978-1-58113-216-8. DOI: 10.1145/332040.332433. URL <https://dl.acm.org/doi/10.1145/332040.332433>.
- Bertram Gawronski. Theory-based bias correction in dispositional inference: The fundamental attribution error is dead, long live the correspondence bias. *European Review of Social Psychology*, 15(1):183–217, January 2004. ISSN 1046-3283. DOI: 10.1080/10463280440000026. URL <https://doi.org/10.1080/10463280440000026>.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for Datasets, December 2021. URL <http://arxiv.org/abs/1803.09010>. arXiv:1803.09010 [cs].
- Julie Gerlings and Ioanna Constantiou. Machine Learning in Transaction Monitoring: The Prospect of xAI, December 2022. URL <http://arxiv.org/abs/2210.07648>. arXiv:2210.07648 [cs].
- Zoubin Ghahramani. Unsupervised Learning. In Olivier Bousquet, Ulrike von Luxburg, and Gunnar Rätsch, editors, *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Revised Lectures*, Lecture Notes in Computer Science, pages 72–112. Springer, Berlin, Heidelberg, 2004. ISBN 978-3-540-28650-9. URL https://doi.org/10.1007/978-3-540-28650-9_5.

- Bhavya Ghai, Q. Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. Explainable Active Learning (XAL): Toward AI Explanations as Interfaces for Machine Teachers. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–28, January 2021. ISSN 2573-0142. DOI: 10.1145/3432934. URL <https://dl.acm.org/doi/10.1145/3432934>.
- Asma Ghandeharioun, Been Kim, Chun-Liang Li, Brendan Jou, Brian Eoff, and Rosalind W. Picard. DISSECT: Disentangled Simultaneous Explanations via Concept Traversals. *arXiv:2105.15164 [cs]*, February 2022. URL <http://arxiv.org/abs/2105.15164>. arXiv: 2105.15164.
- Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750, November 2021. ISSN 2589-7500. DOI: 10.1016/S2589-7500(21)00208-9. URL <https://www.sciencedirect.com/science/article/pii/S2589750021002089>.
- Azin Ghazimatin, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. ELIXIR: Learning from User Feedback on Explanations to Improve Recommender Models. In *Proceedings of the Web Conference 2021*, WWW '21, pages 3850–3860, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8312-7. DOI: 10.1145/3442381.3449848. URL <https://doi.org/10.1145/3442381.3449848>.
- Gerd Gigerenzer. *The Intelligence of Intuition*. Cambridge University Press, October 2023. ISBN 978-1-00-930490-0. Google-Books-ID: 7IHZEAAAQBAJ.
- L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89, Turin, Italy, October 2018. IEEE. DOI: 10.1109/DSAA.2018.00018.
- Ella Glikson and Anita Williams Woolley. Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals*, 14(2):627–660, July 2020. ISSN 1941-6520. DOI: 10.5465/annals.2018.0057. URL <https://journals.aom.org/doi/10.5465/annals.2018.0057>. Publisher: Academy of Management.
- Dale L. Goodhue and Ronald L. Thompson. Task-Technology Fit and Individual Performance. *MIS Quarterly*, 19(2):213–236, 1995. ISSN 0276-7783. DOI: 10.2307/249689. URL <https://www.jstor.org/stable/249689>. Publisher: Management Information Systems Research Center, University of Minnesota.
- Dimitrios Goranitis and Meral Cailali. Global fines for AML/CFT related issues increase in 2022. Technical report, Deloitte, February 2023.
- Maartje M. A. de Graaf and Bertram F. Malle. How People Explain Action (and Autonomous Intelligent Systems Should Too). In *2017 AAAI Fall Symposium Series*, page 8, Arlington, Virginia, October 2017. AAAI. URL <https://www.aaai.org/ocs/index.php/FSS/FSS17/paper/view/16009>.
- Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. The Dark (Patterns) Side of UX Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–14, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 978-1-4503-5620-6. DOI: 10.1145/3173574.3174108. URL <https://dl.acm.org/doi/10.1145/3173574.3174108>.
- Ben Green and Yiling Chen. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):50:1–50:24, November 2019. DOI: 10.1145/3359152. URL <https://doi.org/10.1145/3359152>.
- H. P. Grice. *Logic and Conversation*. Brill, December 1975. ISBN 978-90-04-36881-1. URL <https://brill.com/view/book/edcoll/9789004368811/BP000003.xml>. Pages: 41-58 Section: Speech Acts.
- G. Mark Grimes, Ryan M. Schuetzler, and Justin Scott Giboney. Mental models and expectation violations in conversational AI interactions. *Decision Support Systems*, 144:113515, May 2021. ISSN 0167-9236. DOI: 10.1016/j.dss.2021.113515. URL <https://www.sciencedirect.com/science/article/pii/S0167923621000257>.

- Rob Gruppetta. Using artificial intelligence to keep criminal funds out of the financial system, December 2017. URL <https://www.fca.org.uk/news/speeches/using-artificial-intelligence-keep-criminal-funds-out-financial-system>.
- Ziwei Gu, Jing Nathan Yan, and Jeffrey M. Rzeszotarski. Understanding User Sensemaking in Machine Learning Fairness Assessment Systems. In *Proceedings of the Web Conference 2021, WWW '21*, pages 658–668, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8312-7. DOI: 10.1145/3442381.3450092. URL <https://doi.org/10.1145/3442381.3450092>.
- Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51(5):93:1–93:42, August 2018. ISSN 0360-0300. DOI: 10.1145/3236009. URL <https://doi.org/10.1145/3236009>.
- David Gunning and David Aha. DARPA’s Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, 40(2):44–58, June 2019. ISSN 2371-9621. DOI: 10.1609/aimag.v40i2.2850. URL <https://www.aaai.org/ojs/index.php/aimagazine/article/view/2850>. Number: 2.
- Neil Gunningham and Darren Sinclair. Organizational Trust and the Limits of Management-Based Regulation. *Law & Society Review*, 43(4):865–900, 2009. ISSN 1540-5893. DOI: 10.1111/j.1540-5893.2009.00391.x. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-5893.2009.00391.x>. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-5893.2009.00391.x>.
- Lijie Guo, Elizabeth M. Daly, Ozgur Alkan, Massimiliano Mattetti, Owen Cornec, and Bart Knijnenburg. Building Trust in Interactive Machine Learning via User Contributed Interpretable Rules. In *27th International Conference on Intelligent User Interfaces, UII '22*, pages 537–548, New York, NY, USA, March 2022. Association for Computing Machinery. ISBN 978-1-4503-9144-3. DOI: 10.1145/3490099.3511111. URL <https://doi.org/10.1145/3490099.3511111>.
- Ruo Cheng Guo, Lu Cheng, Jundong Li, P. Richard Hahn, and Huan Liu. A Survey of Learning Causality with Data: Problems and Methods. *ACM Computing Surveys*, 53(4):1–37, July 2021. ISSN 0360-0300, 1557-7341. DOI: 10.1145/3397269. URL <https://dl.acm.org/doi/10.1145/3397269>.
- Abhishek Gupta, Dwijendra Nath Dwivedi, and Jigar Shah. *Artificial Intelligence Applications in Banking and Financial Services: Anti Money Laundering and Compliance*. Springer Nature, July 2023. ISBN 978-981-9925-71-1. Google-Books-ID: c2LMEAAAQBAJ.
- Christos Hadjiemmanuil. A Heavily Regulated Industry: The Varied Objectives of Financial Regulation, December 2015. URL <https://papers.ssrn.com/abstract=2733062>.
- Joseph Y. Halpern and Judea Pearl. Causes and Explanations: A Structural-Model Approach. Part I: Causes. *The British Journal for the Philosophy of Science*, 56(4):843–887, December 2005. ISSN 0007-0882. DOI: 10.1093/bjps/axi147. URL <https://www.journals.uchicago.edu/doi/10.1093/bjps/axi147>. Publisher: The University of Chicago Press.
- Ronan Hamon, Henrik Junklewitz, and Ignacio Sanchez. Robustness and Explainability of Artificial Intelligence. JRC Technical Report EUR 30040 EN, European Commission Joint Research Center, 2020.
- Ronan Hamon, Henrik Junklewitz, Ignacio Sanchez, Gianclaudio Malgieri, and Paul De Hert. Bridging the Gap Between AI and Explainability in the GDPR: Towards Trustworthiness-by-Design in Automated Decision-Making. *IEEE Computational Intelligence Magazine*, 17(1):72–85, February 2022. ISSN 1556-6048. DOI: 10.1109/MCI.2021.3129960. Conference Name: IEEE Computational Intelligence Magazine.
- Russell Hardin. *Trust*. Polity, April 2006. ISBN 978-0-7456-2465-5. Google-Books-ID: XWwpRhX1tdsC.
- Gilbert H. Harman. The Inference to the Best Explanation. *The Philosophical Review*, 74(1):88–95, 1965. ISSN 0031-8108. DOI: 10.2307/2183532. URL <https://www.jstor.org/stable/2183532>. Publisher: [Duke University Press, Philosophical Review].

- Peter Hedström and Petri Ylikoski. Causal Mechanisms in the Social Sciences. *Annual Review of Sociology*, 36(1):49–67, 2010. DOI: 10.1146/annurev.soc.012809.102632. URL <https://doi.org/10.1146/annurev.soc.012809.102632>. _eprint: <https://doi.org/10.1146/annurev.soc.012809.102632>.
- Clément Henin and Daniel Le Métayer. Beyond explainability: justifiability and contestability of algorithmic decision systems. *AI & SOCIETY*, 37(4):1397–1410, December 2022. ISSN 1435-5655. DOI: 10.1007/s00146-021-01251-8. URL <https://doi.org/10.1007/s00146-021-01251-8>.
- Sam Hepenstal, Leishi Zhang, Neesha Kodagoda, and B. I. William Wong. Developing Conversational Agents for Use in Criminal Investigations. *ACM Transactions on Interactive Intelligent Systems*, 11(3-4): 1–35, December 2021. ISSN 2160-6455, 2160-6463. DOI: 10.1145/3444369. URL <https://dl.acm.org/doi/10.1145/3444369>.
- Bernee Herman. The Promise and Peril of Human Evaluation for Model Interpretability. *arXiv:1711.07414 [cs, stat]*, October 2019. URL <http://arxiv.org/abs/1711.07414>. arXiv: 1711.07414.
- Diana C. Hernandez-Bocanegra and Jürgen Ziegler. Conversational review-based explanations for recommender systems: Exploring users’ query behavior. In *CUI 2021 - 3rd Conference on Conversational User Interfaces*, CUI ’21, pages 1–11, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8998-3. DOI: 10.1145/3469595.3469596. URL <https://doi.org/10.1145/3469595.3469596>.
- Scarlett R. Herring, Chia-Chen Chang, Jesse Krantzler, and Brian P. Bailey. Getting inspired! understanding how and why examples are used in creative design practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’09, pages 87–96, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 978-1-60558-246-7. DOI: 10.1145/1518701.1518717. URL <https://doi.org/10.1145/1518701.1518717>.
- Christian Herzog. On the risk of confusing interpretability with explicability. *AI and Ethics*, 2(1):219–225, February 2022. ISSN 2730-5961. DOI: 10.1007/s43681-021-00121-9. URL <https://doi.org/10.1007/s43681-021-00121-9>.
- Germund Hesslow. The Problem of Causal Selection. In Denis J. Hilton, editor, *Contemporary Science and Natural Explanation: Commonsense Conceptions of Causality*. New York University Press, 1988.
- High-Level Expert Group on AI (HLEG). A definition of AI: Main Capabilities and Scientific Disciplines. Technical report, European Commission, Brussels, December 2018.
- High-Level Expert Group on AI (HLEG). Ethics guidelines for trustworthy AI | Shaping Europe’s digital future. Technical report, European Commission, April 2019. URL <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>.
- Mireille Hildebrandt. Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning. *Theoretical Inquiries in Law*, 20(1):83–121, January 2019. ISSN 1565-3404. DOI: 10.1515/til-2019-0004. URL <https://www.degruyter.com/document/doi/10.1515/til-2019-0004/html>. Publisher: De Gruyter.
- Denis J. Hilton. Logic and causal attribution. In *Contemporary science and natural explanation: Commonsense conceptions of causality*, pages 33–65. New York University Press, New York, NY, US, 1988. ISBN 978-0-8147-3443-8.
- Denis J. Hilton and Ben R. Slugoski. Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, 93(1):75–88, 1986. ISSN 1939-1471. DOI: 10.1037/0033-295X.93.1.75. Place: US Publisher: American Psychological Association.
- Michael Hind. Explaining explainable AI. *XRDS: Crossroads, The ACM Magazine for Students*, 25(3):16–19, 2019. ISSN 1528-4972. DOI: 10.1145/3313096. URL <https://doi.org/10.1145/3313096>.
- Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, November 1997. ISSN 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL <https://ieeexplore.ieee.org/abstract/document/6795963>. Conference Name: Neural Computation.

- Marit Hoegen, Hilko van Rooijen, and Maarten Rijssenbeek. Three fundamental changes to the Dutch AML system, 2023. URL <https://www2.deloitte.com/nl/nl/pages/finance/articles/three-fundamental-changes-to-the-dutch-aml-system.html>.
- Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Metrics for Explainable AI: Challenges and Prospects, February 2019. URL <http://arxiv.org/abs/1812.04608>. arXiv:1812.04608 [cs].
- Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Glasgow Scotland Uk, May 2019. ACM. ISBN 978-1-4503-5970-2. DOI: 10.1145/3290605.3300809. URL <https://dl.acm.org/doi/10.1145/3290605.3300809>.
- Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, 9(4): e1312, 2019. ISSN 1942-4795. DOI: 10.1002/widm.1312. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1312>. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1312>.
- Andreas Holzinger, André Carrington, and Heimo Müller. Measuring the Quality of Explanations: The System Causability Scale (SCS). *KI - Künstliche Intelligenz*, 34(2):193–198, June 2020. ISSN 1610-1987. DOI: 10.1007/s13218-020-00636-z. URL <https://doi.org/10.1007/s13218-020-00636-z>.
- Andreas Holzinger, Bernd Malle, Anna Saranti, and Bastian Pfeifer. Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI. *Information Fusion*, 71: 28–37, July 2021. ISSN 1566-2535. DOI: 10.1016/j.inffus.2021.01.008. URL <https://www.sciencedirect.com/science/article/pii/S1566253521000142>.
- Samuel Huron. *Constructive Visualization : A token-based paradigm allowing to assemble dynamic visual representation for non-experts*. These de doctorat, Paris 11, September 2014. URL <https://www.theses.fr/2014PA112253>.
- Johan Huysmans, Karel Dejaeger, Christophe Mues, Jan Vanthienen, and Bart Baesens. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 51(1):141–154, April 2011. ISSN 0167-9236. DOI: 10.1016/j.dss.2010.12.003. URL <https://www.sciencedirect.com/science/article/pii/S0167923610002368>.
- International Organization for Standardization (ISO). Ergonomics of human-system interaction Human-centred design for interactive systems.
- International Organization for Standardization (ISO). Artificial Overview of trustworthiness in artificial intelligence, January 2022. URL <https://www.iso.org/standard/77608.html>.
- Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C. Ahn, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, Yokohama Japan, May 2021. ACM. ISBN 978-1-4503-8096-6. DOI: 10.1145/3411764.3445385. URL <https://dl.acm.org/doi/10.1145/3411764.3445385>.
- Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, pages 624–635, New York, NY, USA, March 2021. Association for Computing Machinery. ISBN 978-1-4503-8309-7. DOI: 10.1145/3442188.3445923. URL <https://dl.acm.org/doi/10.1145/3442188.3445923>.
- Lars-Erik Janlert and Erik Stolterman. The Meaning of Interactivity—Some Proposals for Definitions and Measures. *Human-Computer Interaction*, 32(3):103–138, May 2017. ISSN 0737-0024. DOI: 10.1080/07370024.2016.1226139. URL <https://doi.org/10.1080/07370024.2016.1226139>. Publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/07370024.2016.1226139>.

- Shichao Jia, Zeyu Li, Nuo Chen, and Jiawan Zhang. Towards Visual Explainable Active Learning for Zero-Shot Classification. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):791–801, January 2022. ISSN 1941-0506. DOI: 10.1109/TVCG.2021.3114793.
- Zhuochen Jin, Shuyuan Cui, Shunan Guo, David Gotz, Jimeng Sun, and Nan Cao. CarePre: An Intelligent Clinical Decision Assistance System. *ACM Transactions on Computing for Healthcare*, 1(1):6:1–6:20, March 2020. ISSN 2691-1957. DOI: 10.1145/3344258. URL <https://doi.org/10.1145/3344258>.
- Joanna Bryson. AI & Global Governance: No One Should Trust AI, November 2018. URL <https://unu.edu/cpr/blog-post/ai-global-governance-no-one-should-trust-ai>. Accessed 1/22/2024.
- Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, September 2019. ISSN 2522-5839. DOI: 10.1038/s42256-019-0088-2. URL <https://www.nature.com/articles/s42256-019-0088-2>. Number: 9 Publisher: Nature Publishing Group.
- Patrick W. Jordan, B. Thomas, Ian Lyall McClelland, and Bernard Weerdmeester. *Usability Evaluation In Industry*. CRC Press, June 1996. ISBN 978-1-4987-1041-1. Google-Books-ID: ujFRDwAAQBAJ.
- M. Jullum, A. Løland, R.B. Huseby, G. Ånonsen, and J. Lorentzen. Detecting money laundering transactions with machine learning. *Journal of Money Laundering Control*, 23(1):173–186, 2020. DOI: 10.1108/JMLC-07-2019-0055.
- Daniel Kahneman. *Thinking, fast and slow*. Thinking, fast and slow. Farrar, Straus and Giroux, New York, NY, US, 2011. ISBN 978-0-374-27563-1 978-1-4299-6935-2. Pages: 499.
- Daniel Kahneman and Gary Klein. Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, 64(6):515–526, 2009. ISSN 1935-990X. DOI: 10.1037/a0016755. Place: US Publisher: American Psychological Association.
- Daniel Kahneman and Amos Tversky. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2):263–291, 1979. ISSN 0012-9682. DOI: 10.2307/1914185. URL <https://www.jstor.org/stable/1914185>. Publisher: [Wiley, Econometric Society].
- Daniel Kahneman, Stewart Paul Slovic, Paul Slovic, Amos Tversky, and Cambridge University Press. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, April 1982. ISBN 978-0-521-28414-1. Google-Books-ID: _oH8gwj4a1MC.
- Margot E. Kaminski and Jennifer M. Urban. The Right to Contest AI, November 2021. URL <https://papers.ssrn.com/abstract=3965041>.
- Katrina Zhu. The State of State AI Laws: 2023, August 2023. URL <https://epic.org/the-state-of-state-ai-laws-2023/>.
- Davinder Kaur, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durresi. Trustworthy Artificial Intelligence: A Review. *ACM Computing Surveys*, 55(2):39:1–39:38, January 2022. ISSN 0360-0300. DOI: 10.1145/3491209. URL <https://dl.acm.org/doi/10.1145/3491209>.
- Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting Interpretability: Understanding Data Scientists’ Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, New York, NY, USA, April 2020. Association for Computing Machinery. ISBN 978-1-4503-6708-0. URL <https://doi.org/10.1145/3313831.3376219>.
- D.A. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, January 2002. ISSN 1941-0506. DOI: 10.1109/2945.981847. Conference Name: IEEE Transactions on Visualization and Computer Graphics.
- Carmel Kent, Esther Laslo, and Sheizaf Rafaeli. Interactivity in online discussions and learning outcomes. *Computers & Education*, 97:116–128, June 2016. ISSN 0360-1315. DOI: 10.1016/j.compedu.2016.03.002. URL <https://www.sciencedirect.com/science/article/pii/S0360131516300537>.

- Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T. Hancock, and Michael S. Bernstein. Conceptual Metaphors Impact Perceptions of Human-AI Collaboration. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):163:1–163:26, October 2020. DOI: 10.1145/3415234. URL <https://dl.acm.org/doi/10.1145/3415234>.
- Anjali Khurana, Parsa Alamzadeh, and Parmit K. Chilana. ChatrEx: Designing Explainable Chatbot Interfaces for Enhancing Usefulness, Transparency, and Trust. In *2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 1–11, October 2021. DOI: 10.1109/VL/HCC51201.2021.9576440. ISSN: 1943-6106.
- Been Kim, Elena Glassman, Brittney Johnson, and Julie Shah. iBCM: Interactive Bayesian Case Model Empowering Humans via Intuitive Interaction. April 2015. URL <https://dspace.mit.edu/handle/1721.1/96315>. Accepted: 2015-04-01T17:30:03Z.
- Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! Criticism for Interpretability. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL <https://papers.nips.cc/paper/2016/hash/5680522b8e2bb01943234bce7bf84534-Abstract.html>.
- Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, pages 2668–2677. PMLR, July 2018. URL <https://proceedings.mlr.press/v80/kim18d.html>. ISSN: 2640-3498.
- Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–17, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 978-1-4503-9421-5. DOI: 10.1145/3544548.3581001. URL <https://dl.acm.org/doi/10.1145/3544548.3581001>.
- Taenyun Kim and Hayeon Song. The Effect of Message Framing and Timing on the Acceptance of Artificial Intelligence's Suggestion. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, pages 1–8, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 978-1-4503-6819-3. DOI: 10.1145/3334480.3383038. URL <https://doi.org/10.1145/3334480.3383038>.
- Kim, Chris, Lin, Xiao, Collins, Christopher, Taylor, Graham W, and Amer, Mohamed R. Learn, Generate, Rank, Explain: A Case Study of Visual Explanation by Generative Machine Learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, August 2021. DOI: 10.1145/3465407. URL <https://dl.acm.org/doi/abs/10.1145/3465407>. Publisher: ACM PUB27 New York, NY.
- Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (Un)reliability of saliency methods, November 2017. URL <http://arxiv.org/abs/1711.00867>. arXiv:1711.00867 [cs, stat].
- Alexandra Kirsch. Explain to whom? Putting the User in the Center of Explainable AI. In *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017 co-located with 16th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2017)*, Bari, Italy, 2017. URL <https://hal.archives-ouvertes.fr/hal-01845135>.
- René F. Kizilcec. 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*, CHI '16, pages 2390–2395, New York, NY, USA, May 2016. Association for Computing Machinery. ISBN 978-1-4503-3362-7. DOI: 10.1145/2858036.2858402. URL <https://doi.org/10.1145/2858036.2858402>.
- Paul A. Klaczynski, David H. Gordon, and James Fauth. Goal-oriented critical reasoning and individual differences in critical reasoning biases. *Journal of Educational Psychology*, 89(3):470–485, 1997. ISSN 1939-2176. DOI: 10.1037/0022-0663.89.3.470. Place: US Publisher: American Psychological Association.

- Gary A. Klein. *Sources of Power: How People Make Decisions*. Nature, 1988. ISBN 978-0-262-53429-1. Google-Books-ID: JW01DwAAQBAJ.
- Tomáš Kliegr, Štěpán Bahník, and Johannes Fürnkranz. A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence*, 295:103458, June 2021. ISSN 0004-3702. DOI: 10.1016/j.artint.2021.103458. URL <https://www.sciencedirect.com/science/article/pii/S0004370221000096>.
- Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4):441–504, October 2012. ISSN 1573-1391. DOI: 10.1007/s11257-011-9118-4. URL <https://doi.org/10.1007/s11257-011-9118-4>.
- Derek J. Koehler. *Explanation, Imagination, and Confidence in Judgment*. 1991.
- Pang Wei Koh and Percy Liang. Understanding Black-box Predictions via Influence Functions, December 2020. URL <http://arxiv.org/abs/1703.04730>. arXiv:1703.04730 [cs, stat].
- Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020.
- Adriano Koshiyama, Emre Kazim, Philip Treleaven, Pete Rai, Lukasz Szpruch, Giles Pavey, Ghazi Ahamat, Franziska Leutner, Randy Goebel, Andrew Knight, Janet Adams, Christina Hitrova, Jeremy Barnett, Parashkev Nachev, David Barber, Tomas Chamorro-Premuzic, Konstantin Klemmer, Miro Gregorovic, Shakeel Khan, and Elizabeth Lomas. Towards Algorithm Auditing: A Survey on Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms. *SSRN Electronic Journal*, page 31, 2021. ISSN 1556-5068. DOI: 10.2139/ssrn.3778998. URL <https://www.ssrn.com/abstract=3778998>.
- Yubo Kou and Xinning Gui. Mediating Community-AI Interaction through Situated Explanation: The Case of AI-Led Moderation. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):102:1–102:27, October 2020. DOI: 10.1145/3415173. URL <https://doi.org/10.1145/3415173>.
- Pigi Kouki, James Schaffer, Jay Pujara, John O’Donovan, and Lise Getoor. Personalized explanations for hybrid recommender systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI ’19*, pages 379–390, New York, NY, USA, March 2019. Association for Computing Machinery. ISBN 978-1-4503-6272-6. DOI: 10.1145/3301275.3302306. URL <https://doi.org/10.1145/3301275.3302306>.
- Maria Kouvela, Ilias Dimitriadis, and Athena Vakali. Bot-Detective: An explainable Twitter bot detection service with crowdsourcing functionalities. In *Proceedings of the 12th International Conference on Management of Digital EcoSystems, MEDES ’20*, pages 55–63, New York, NY, USA, November 2020. Association for Computing Machinery. ISBN 978-1-4503-8115-4. DOI: 10.1145/3415958.3433075. URL <https://doi.org/10.1145/3415958.3433075>.
- Josua Krause, Adam Perer, and Kenney Ng. Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI ’16*, pages 5686–5697, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 978-1-4503-3362-7. DOI: 10.1145/2858036.2858529. URL <https://doi.org/10.1145/2858036.2858529>.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL <https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html>.
- Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. Accountable Algorithms, March 2016. URL <https://papers.ssrn.com/abstract=2765268>.

- Luisa Kruse, Nico Wunderlich, and Roman Beck. Artificial Intelligence for the Financial Services Industry: What Challenges Organizations to Succeed. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, page 10, Hawaii, January 2019. ScholarSpace. ISBN 978-0-9981331-2-6. URL <http://hdl.handle.net/10125/60075>.
- Ouren Kuiper, Martin van den Berg, Joost van der Burgt, and Stefan Leijnen. Exploring explainable AI in the financial sector: Perspectives of banks and supervisory authorities. In *Artificial Intelligence and Machine Learning: 33rd Benelux Conference on Artificial Intelligence*, pages 105–119, Esch-sur-Alzette, Luxembourg, November 2021. Springer.
- T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W. Wong. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, pages 3–10, September 2013. DOI: 10.1109/VLHCC.2013.6645235. ISSN: 1943-6106.
- Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces, IUI '15*, pages 126–137, New York, NY, USA, March 2015. Association for Computing Machinery. ISBN 978-1-4503-3306-1. DOI: 10.1145/2678025.2701399. URL <https://doi.org/10.1145/2678025.2701399>.
- I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with Shapley-value-based explanations as feature importance measures. *arXiv:2002.11097 [cs, stat]*, June 2020. URL <http://arxiv.org/abs/2002.11097>. arXiv: 2002.11097.
- E. Kurshan and H. Shen. Graph Computing for Financial Crime and Fraud Detection: Trends, Challenges and Outlook, March 2021. URL <http://arxiv.org/abs/2103.03227>. arXiv:2103.03227 [cs].
- Alexander Kurz, Katja Hauser, Hendrik Alexander Mehrtens, Eva Krieghoff-Henning, Achim Hekler, Jakob Nikolas Kather, Stefan Fröhling, Christof von Kalle, and Titus Josef Brinker. Uncertainty Estimation in Medical Image Classification: Systematic Review. *JMIR Medical Informatics*, 10(8):e36427, August 2022. DOI: 10.2196/36427. URL <https://medinform.jmir.org/2022/8/e36427>. Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada.
- Dattatray Vishnu Kute, Biswajeet Pradhan, Nagesh Shukla, and Abdullah Alamri. Deep Learning and Explainable Artificial Intelligence Techniques Applied for Detecting Money Laundering—A Critical Review. *IEEE Access*, 9:82300–82317, 2021. ISSN 2169-3536. DOI: 10.1109/ACCESS.2021.3086230. Conference Name: IEEE Access.
- Bum Chul Kwon, Min-Je Choi, Joanne Taery Kim, Edward Choi, Young Bin Kim, Soonwook Kwon, Jimeng Sun, and Jaegul Choo. RetainVis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):299–309, January 2019. ISSN 1941-0506. DOI: 10.1109/TVCG.2018.2865027. Conference Name: IEEE Transactions on Visualization and Computer Graphics.
- Nevine Makram Labib, Mohammed Abo Rizka, and Amr Ehab Muhammed Shokry. Survey of Machine Learning Approaches of Anti-money Laundering Techniques to Counter Terrorism Finance. In Atef Zaki Ghalwash, Nashaat El Khameesy, Dalia A. Magdi, and Amit Joshi, editors, *Internet of Things—Applications and Future*, Lecture Notes in Networks and Systems, pages 73–87, Singapore, 2020. Springer. ISBN 9789811530753.
- Vivian Lai and Chenhao Tan. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19*, pages 29–38, New York, NY, USA, January 2019. Association for Computing Machinery. ISBN 978-1-4503-6125-5. DOI: 10.1145/3287560.3287590. URL <https://doi.org/10.1145/3287560.3287590>.

- Vivian Lai, Chacha Chen, Q. Vera Liao, Alison Smith-Renner, and Chenhao Tan. Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies, December 2021. URL <http://arxiv.org/abs/2112.11471>. arXiv:2112.11471 [cs].
- Himabindu Lakkaraju and Osbert Bastani. "How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 79–85, 2020. URL <https://doi.org/10.1145/3375627.3375833>.
- Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Interpretable & Explorable Approximations of Black Box Models. *arXiv:1707.01154 [cs]*, July 2017. URL <http://arxiv.org/abs/1707.01154>. arXiv: 1707.01154.
- M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sasing, and K. Baum. What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, 296, 2021. DOI: 10.1016/j.artint.2021.103473.
- Matthias Laporte. ACPR Conference, p.85, "LUCIA": a SupTech tool to support the fight against money laundering and terrorism financing, November 2021. URL https://acpr.banque-france.fr/sites/default/files/media/2022/11/15/20211126_presentations_des_intervenants_de_la_matinée.pdf.
- David B. Leake. Goal-based explanation evaluation. *Cognitive Science*, 15(4):509–545, October 1991. ISSN 0364-0213. DOI: 10.1016/0364-0213(91)80017-Y. URL <https://www.sciencedirect.com/science/article/pii/036402139180017Y>.
- David B. Leake. Abduction, experience, and goals: a model of everyday abductive explanation. *Journal of Experimental & Theoretical Artificial Intelligence*, 7(4):407–428, October 1995. ISSN 0952-813X. DOI: 10.1080/09528139508953820. URL <https://doi.org/10.1080/09528139508953820>. Publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/09528139508953820>.
- Matthew L. Leavitt and Ari Morcos. Towards falsifiable interpretability research. *arXiv:2010.12016 [cs, stat]*, October 2020. URL <http://arxiv.org/abs/2010.12016>. arXiv: 2010.12016.
- Freddy Lecue. On the role of knowledge graphs in explainable AI. *Semantic Web*, 11(1):41–51, January 2020. ISSN 1570-0844. DOI: 10.3233/SW-190374. URL <https://content.iospress.com/articles/semantic-web/sw190374>. Publisher: IOS Press.
- Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. ISSN 1476-4687. DOI: 10.1038/nature14539. URL <https://www.nature.com/articles/nature14539>. Number: 7553 Publisher: Nature Publishing Group.
- John Lee and Neville Moray. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10):1243–1270, October 1992. ISSN 0014-0139. DOI: 10.1080/00140139208967392. URL <https://doi.org/10.1080/00140139208967392>. Publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/00140139208967392>.
- John D. Lee and Katrina A. See. Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, 46(1):50–80, March 2004. ISSN 0018-7208. URL https://journals.sagepub.com/doi/abs/10.1518/hfes.46.1.50_30392. Publisher: SAGE Publications Inc.
- Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):182:1–182:26, November 2019. DOI: 10.1145/3359284. URL <https://doi.org/10.1145/3359284>.
- Randy Lee. Louis Brandeis's Vision of Light and Justice as Articulated on the Side of Coffee Mug. *Touro Law Review*, 33:323, 2017. URL <https://heinonline.org/HOL/Page?handle=hein.journals/touro33&id=331&div=&collection=>.

- Michael Levi and Peter Reuter. Money Laundering. *Crime and Justice*, 34:289–375, January 2006. ISSN 0192-3234. DOI: 10.1086/501508. URL <https://www.journals.uchicago.edu/doi/abs/10.1086/501508>. Publisher: The University of Chicago Press.
- David Levi-Faur. *Handbook on the Politics of Regulation*. Edward Elgar Publishing, January 2011. ISBN 978-0-85793-611-0. Google-Books-ID: KOKtKzEyQIYC.
- Michael E. Levine and Jennifer L. Forrence. Regulatory Capture, Public Interest, and the Public Agenda: Toward a Synthesis. *Journal of Law, Economics, and Organization*, 6:167, 1990. URL <https://heinonline.org/HOL/Page?handle=hein.journals/jleo6&id=651&div=&collection=>. Accessed 11/29/2023.
- James R. Lewis. Psychometric evaluation of an after-scenario questionnaire for computer usability studies: the ASQ. *ACM SIGCHI Bulletin*, 23(1):78–81, January 1991. ISSN 0736-6906. DOI: 10.1145/122672.122692. URL <https://doi.org/10.1145/122672.122692>.
- Q. Vera Liao and Kush R. Varshney. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences, April 2022. URL <http://arxiv.org/abs/2110.10790>. arXiv:2110.10790 [cs].
- Q. Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–15, New York, NY, USA, April 2020. Association for Computing Machinery. ISBN 978-1-4503-6708-0. DOI: 10.1145/3313831.3376590. URL <https://doi.org/10.1145/3313831.3376590>.
- Q. Vera Liao, Hariharan Subramonyam, Jennifer Wang, and Jennifer Wortman Vaughan. Designerly Understanding: Information Needs for Model Transparency to Support Design Ideation for AI-Powered User Experience. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21, Hamburg Germany, April 2023. ACM. ISBN 978-1-4503-9421-5. DOI: 10.1145/3544548.3580652. URL <https://dl.acm.org/doi/10.1145/3544548.3580652>.
- Brian Y. Lim and Anind K. Dey. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th international conference on Ubiquitous computing*, UbiComp '09, pages 195–204, New York, NY, USA, September 2009. Association for Computing Machinery. ISBN 978-1-60558-431-7. DOI: 10.1145/1620545.1620576. URL <https://doi.org/10.1145/1620545.1620576>.
- Gabriel Lima, Nina Grgić-Hlača, Jin Keun Jeong, and Meeyoung Cha. The Conflict Between Explainable and Accountable Decision-Making Algorithms. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 2103–2113, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 978-1-4503-9352-2. DOI: 10.1145/3531146.3534628. URL <https://dl.acm.org/doi/10.1145/3531146.3534628>.
- Adam Dahlgren Lindström, Wendy E. Mackay, and Virginia Dignum. Thinking Fast And Slow In Human-Centered AI. In *Thinking Fast and Slow and Other Cognitive Theories in AI, AAAI Fall symposium FSS-22*, pages 3–pages, 2022. URL <https://inria.hal.science/hal-03991946/document>.
- Walter Lippmann. *The Phantom Public*. Transaction Publishers, 1993. ISBN 978-1-56000-677-0. Google-Books-ID: AUJTAQAAQBAJ.
- Peter Lipton. Contrastive Explanation. *Royal Institute of Philosophy Supplements*, 27:247–266, March 1990. ISSN 1755-3555, 1358-2461. Publisher: Cambridge University Press.
- Zachary C. Lipton. The mythos of model interpretability. *Communications of the ACM*, 61(10):36–43, September 2018. ISSN 0001-0782. DOI: 10.1145/3233231. URL <https://doi.org/10.1145/3233231>.
- Han Liu, Vivian Lai, and Chenhao Tan. Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):408:1–408:45, October 2021. DOI: 10.1145/3479552. URL <https://doi.org/10.1145/3479552>.

- Jiali Liu. *Data expression : understanding and supporting alternatives in data analysis processes*. phdthesis, Institut Polytechnique de Paris, September 2021. URL <https://theses.hal.science/tel-03577013>.
- Tania Lombrozo. The structure and function of explanations. *Trends in Cognitive Sciences*, 10(10):464–470, October 2006. ISSN 1364-6613. DOI: 10.1016/j.tics.2006.08.004. URL <http://www.sciencedirect.com/science/article/pii/S1364661306002117>.
- Tania Lombrozo. Simplicity and probability in causal explanation. *Cognitive Psychology*, 55(3):232–257, November 2007. ISSN 0010-0285. DOI: 10.1016/j.cogpsych.2006.09.006. URL <https://www.sciencedirect.com/science/article/pii/S0010028506000739>.
- Tania Lombrozo. Explanatory Preferences Shape Learning and Inference. *Trends in Cognitive Sciences*, 20(10):748–759, October 2016. ISSN 1364-6613. DOI: 10.1016/j.tics.2016.08.001. URL <https://www.sciencedirect.com/science/article/pii/S136466131630105X>.
- Tanya Lombrozo. Explanation and Abductive Inference. In Keith J. Holyoak and Robert G. Morrison, editors, *The Oxford Handbook of Thinking and Reasoning*, page 0. Oxford University Press, March 2012. ISBN 978-0-19-973468-9. DOI: 10.1093/oxfordhb/9780199734689.013.0014. URL <https://doi.org/10.1093/oxfordhb/9780199734689.013.0014>.
- Luca Longo, Randy Goebel, Freddy Lecue, Peter Kieseberg, and Andreas Holzinger. Explainable Artificial Intelligence: Concepts, Applications, Research Challenges and Visions. In Andreas Holzinger, Peter Kieseberg, A Min Tjoa, and Edgar Weippl, editors, *Machine Learning and Knowledge Extraction*, Lecture Notes in Computer Science, pages 1–16, Cham, 2020. Springer International Publishing. ISBN 978-3-030-57321-8.
- Joana Lorenz, Maria Inês Silva, David Aparício, João Tiago Ascensão, and Pedro Bizarro. Machine learning methods to detect money laundering in the Bitcoin blockchain in the presence of label scarcity. *arXiv:2005.14635 [cs, stat]*, May 2020. URL <http://arxiv.org/abs/2005.14635>. arXiv: 2005.14635 version: 1.
- Jordan J Louviere, Terry N Flynn, and Richard T Carson. Discrete Choice Experiments Are Not Conjoint Analysis. *Journal of Choice Modelling*, 3(3):57–72, January 2010. ISSN 1755-5345. DOI: 10.1016/S1755-5345(13)70014-9. URL <https://www.sciencedirect.com/science/article/pii/S1755534513700149>.
- Ana Lucic, Hinda Haned, and Maarten de Rijke. Why does my model fail? contrastive local explanations for retail forecasting. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 90–98, New York, NY, USA, January 2020. Association for Computing Machinery. ISBN 978-1-4503-6936-7. DOI: 10.1145/3351095.3372824. URL <https://doi.org/10.1145/3351095.3372824>.
- Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 978-1-5108-6096-4.
- Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. Consistent Individualized Feature Attribution for Tree Ensembles, March 2019. URL <http://arxiv.org/abs/1802.03888>. arXiv:1802.03888 [cs, stat].
- Michal Luria. Co-Design Perspectives on Algorithm Transparency Reporting: Guidelines and Prototypes. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1076–1087, Chicago IL USA, June 2023. ACM. ISBN 9798400701924. DOI: 10.1145/3593013.3594064. URL <https://dl.acm.org/doi/10.1145/3593013.3594064>.
- Henrietta Lyons, Eduardo Velloso, and Tim Miller. Designing for Contestation: Insights from Administrative Law. *arXiv:2102.04559 [cs]*, February 2021. URL <http://arxiv.org/abs/2102.04559>. arXiv: 2102.04559.
- Légifrance. Arrêté du 3 novembre 2014 relatif au contrôle interne des entreprises du secteur de la banque, des services de paiement et des services d'investissement soumises au contrôle de l'Autorité de contrôle prudentiel et de résolution, August 2023a. URL <https://www.legifrance.gouv.fr/loda/id/JORFTEXT000029700770>.

- Légifrance. Chapitre Ier : Obligations relatives à la lutte contre le blanchiment des capitaux et le financement du terrorisme (Articles L561-1 à L561-50), August 2023b. URL https://www.legifrance.gouv.fr/codes/section_lc/LEGITEXT000006072026/LEGISCTA000006154830/.
- Wendy E. Mackay and Anne-Laure Fayard. HCI, natural science and design: a framework for triangulation across disciplines. In *Proceedings of the 2nd conference on Designing interactive systems: processes, practices, methods, and techniques*, DIS '97, pages 223–234, New York, NY, USA, 1997. Association for Computing Machinery. ISBN 978-0-89791-863-3. DOI: 10.1145/263552.263612. URL <https://dl.acm.org/doi/10.1145/263552.263612>.
- Poornima Madhavan, Douglas A. Wiegmann, and Frank C. Lacson. Automation Failures on Tasks Easily Performed by Operators Undermine Trust in Automated Aids. *Human Factors*, 48(2):241–256, June 2006. ISSN 0018-7208. DOI: 10.1518/00187200677724408. URL <https://doi.org/10.1518/00187200677724408>. Publisher: SAGE Publications Inc.
- P. Madumal, L. Sonenberg, T. Miller, and F. Vetere. A grounded interaction protocol for explainable artificial intelligence. volume 2, pages 1033–1041, 2019.
- Bertram F. Malle. *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction*. A Bradford Book, Cambridge, MA, USA, September 2004. ISBN 978-0-262-13445-3.
- Nicholas Maltbie, Nan Niu, Matthew Van Doren, and Reese Johnson. XAI tools in the public sector: a case study on predicting combined sewer overflows. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2021*, pages 1032–1044, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8562-6. DOI: 10.1145/3468264.3468547. URL <https://doi.org/10.1145/3468264.3468547>.
- YOKOI-ARAI Mamiko. The Impact of Big Data and Artificial Intelligence (AI) in the Insurance Sector. Technical report, OECD, January 2020. URL <http://www.oecd.org/finance/Impact-Big-Data-AI-in-the-Insurance-Sector.htm>.
- Marisa Tschopp. Digital transformation - Three wrong questions about trust and AI, September 2020. URL <https://digital-commerce.post.ch/en/pages/blog/2020/trust-in-artificial-intelligence>. Accessed 1/11/2024.
- Aniek F. Markus, Jan A. Kors, and Peter R. Rijnbeek. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113:103655, January 2021. ISSN 1532-0464. DOI: 10.1016/j.jbi.2020.103655. URL <https://www.sciencedirect.com/science/article/pii/S1532046420302835>.
- David Martens, Camille Dams, James Hinns, and Mark Vergouwen. Tell Me a Story! Narrative-Driven XAI with Large Language Models, September 2023. URL <http://arxiv.org/abs/2309.17057>. arXiv:2309.17057 [cs].
- Jerry L Mashaw. Small things like reasons are put in a jar: reason and legitimacy in the administrative state. *Fordham Law Review*, 70(1), 2001.
- Arunesh Mathur, Gunes Acar, Michael J. Friedman, Eli Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):81:1–81:32, November 2019. DOI: 10.1145/3359183. URL <https://dl.acm.org/doi/10.1145/3359183>.
- Winston Maxwell. The GDPR and Private Sector Measures to Detect Criminal Activity, March 2021. URL <https://papers.ssrn.com/abstract=3964066>.
- Winston Maxwell. Meaningful Human Control to Detect Algorithmic Errors. In Céline Castets-Renard and Jessica Eynard, editors, *Artificial Intelligence Law: Between Sectoral Rules and Comprehensive Regime - Comparative Law Perspectives*. Bruylant, 2023. URL <https://hal.science/hal-04026883>.

- Winston Maxwell and Bruno Dumas. Meaningful XAI based on user-centric design methodology: Combining legal and human-computer interaction (HCI) approaches to achieve meaningful algorithmic explainability. Technical report, CERRE, 2023.
- Roger C. Mayer, James H. Davis, and F. David Schoorman. An Integrative Model of Organizational Trust. *The Academy of Management Review*, 20(3):709–734, 1995. ISSN 0363-7425. URL <https://www.jstor.org/stable/258792>. Publisher: Academy of Management.
- Elizabeth McCaul. Technology is neither good nor bad, but humans make it so, July 2022. URL <https://www.bankingsupervision.europa.eu/press/speeches/date/2022/html/ssm.sp220713-73f22a486e.en.html>.
- Mike McConville. *Research Methods for Law*. Edinburgh University Press, January 2017. ISBN 978-1-4744-0425-9. Google-Books-ID: 4jRWDwAAQBAJ.
- Ann L. McGill and Jill G. Klein. Contrastive and counterfactual reasoning in causal judgment. *Journal of Personality and Social Psychology*, 64(6):897–905, 1993. ISSN 1939-1315. DOI: 10.1037/0022-3514.64.6.897. Place: US Publisher: American Psychological Association.
- Joseph E. Mcgrath. Methodology Matters: Doing Research in the Behavioral and Social Sciences. In Ronald M. Baecker, Jonathan Grudin, William A.S. Buxton, and Saul Greenberg, editors, *Readings in Human-Computer Interaction*, Interactive Technologies, pages 152–169. Morgan Kaufmann, January 1995. ISBN 978-0-08-051574-8. DOI: 10.1016/B978-0-08-051574-8.50019-4. URL <https://www.sciencedirect.com/science/article/pii/B9780080515748500194>.
- Michael McKenna and D. Justin Coates. Compatibilism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2021 edition, 2021. URL <https://plato.stanford.edu/archives/fall2021/entries/compatibilism/>.
- D. Mcknight, Michelle Carter, Jason Thatcher, and Paul Clay. Trust in a specific technology: An Investigation of its Components and Measures. *ACM Transactions on Management Information Systems*, 2:12–32, June 2011. DOI: 10.1145/1985347.1985353.
- D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. *Information Systems Research*, 13(3):334–359, September 2002. ISSN 1047-7047. DOI: 10.1287/isre.13.3.334.81. URL <https://pubsonline.informs.org/doi/10.1287/isre.13.3.334.81>. Publisher: INFORMS.
- Jessie McWaters and Matthew Blake. Navigating Uncharted Waters: A Roadmap to Responsible Innovation with AI in Financial Services. Part of the Future of Financial Services Series. World Economic Forum. Technical report, World Economic Forum, 2019. URL https://www3.weforum.org/docs/WEF_Navigating_Uncharted_Waters_Report.pdf.
- Gaspar Isaac Melsión, Ilaria Torre, Eva Vidal, and Iolanda Leite. Using Explainability to Help Children Understand Gender Bias in AI. In *Interaction Design and Children*, pages 87–99, Athens Greece, June 2021. ACM. ISBN 978-1-4503-8452-0. DOI: 10.1145/3459990.3460719. URL <https://dl.acm.org/doi/10.1145/3459990.3460719>.
- Bertalan Meskó and Eric J. Topol. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *npj Digital Medicine*, 6(1):1–6, July 2023. ISSN 2398-6352. DOI: 10.1038/s41746-023-00873-0. URL <https://www.nature.com/articles/s41746-023-00873-0>. Number: 1 Publisher: Nature Publishing Group.
- Danaë Metaxa, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, and Christian Sandvig. Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Foundations and Trends® in Human-Computer Interaction*, 14(4):272–344, 2021. ISSN 1551-3955, 1551-3963. DOI: 10.1561/1100000083. URL <http://www.nowpublishers.com/article/Details/HCI-083>.

- Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. To explain or not to explain: the effects of personal characteristics when explaining music recommendations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI '19*, pages 397–407, New York, NY, USA, March 2019. Association for Computing Machinery. ISBN 978-1-4503-6272-6. DOI: 10.1145/3301275.3302313. URL <https://doi.org/10.1145/3301275.3302313>.
- Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, February 2019. ISSN 0004-3702. DOI: 10.1016/j.artint.2018.07.007. URL <http://www.sciencedirect.com/science/article/pii/S0004370218305988>.
- Tim Miller. Contrastive explanation: a structural-model approach. *The Knowledge Engineering Review*, 36:e14, January 2021. ISSN 0269-8889, 1469-8005. DOI: 10.1017/S0269888921000102. URL <https://www.cambridge.org/core/journals/knowledge-engineering-review/article/abs/contrastive-explanation-a-structuralmodel-approach/69A2E32B160C2C7FB65BC88670D7AEA7>. Publisher: Cambridge University Press.
- Tim Miller. Are we measuring trust correctly in explainability, interpretability, and transparency research?, August 2022. URL <http://arxiv.org/abs/2209.00651>. arXiv:2209.00651 [cs].
- Tim Miller. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven decision support, February 2023. URL <http://arxiv.org/abs/2302.12389>. arXiv:2302.12389 [cs].
- Tim Miller, Piers Howe, and Liz Sonenberg. Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. *arXiv:1712.00547 [cs]*, December 2017. URL <http://arxiv.org/abs/1712.00547>. arXiv: 1712.00547.
- Yao Ming, Huamin Qu, and Enrico Bertini. RuleMatrix: Visualizing and Understanding Classifiers with Rules. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):342–352, January 2019. ISSN 1941-0506. DOI: 10.1109/TVCG.2018.2864812. Conference Name: IEEE Transactions on Visualization and Computer Graphics.
- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19*, pages 220–229, New York, NY, USA, January 2019. Association for Computing Machinery. ISBN 978-1-4503-6125-5. DOI: 10.1145/3287560.3287596. URL <https://dl.acm.org/doi/10.1145/3287560.3287596>.
- Melanie Mitchell. Why AI is Harder Than We Think. *arXiv:2104.12871 [cs]*, April 2021. URL <http://arxiv.org/abs/2104.12871>. arXiv: 2104.12871.
- Brent Mittelstadt. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11):501–507, November 2019. ISSN 2522-5839. DOI: 10.1038/s42256-019-0114-4. URL <https://www.nature.com/articles/s42256-019-0114-4>. Number: 11 Publisher: Nature Publishing Group.
- Brent Mittelstadt, Chris Russell, and Sandra Wachter. Explaining Explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19*, pages 279–288, New York, NY, USA, January 2019. Association for Computing Machinery. ISBN 978-1-4503-6125-5. DOI: 10.1145/3287560.3287574. URL <https://doi.org/10.1145/3287560.3287574>.
- Akira Miyake and Priti Shah, editors. *Models of working memory: Mechanisms of active maintenance and executive control*. Models of working memory: Mechanisms of active maintenance and executive control. Cambridge University Press, New York, NY, US, 1999. ISBN 978-0-521-58325-1 978-0-521-58721-1. DOI: 10.1017/CBO9781139174909. Pages: xx, 506.
- David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G. Altman, and The PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine*, 6(7):e1000097, July 2009. ISSN 1549-1676. DOI: 10.1371/journal.pmed.1000097. URL <https://dx.plos.org/10.1371/journal.pmed.1000097>.

- Sina Mohseni, Jeremy E Block, and Eric Ragan. Quantitative Evaluation of Machine Learning Explanations: A Human-Grounded Benchmark. In *26th International Conference on Intelligent User Interfaces*, pages 22–31, College Station TX USA, April 2021a. ACM. ISBN 978-1-4503-8017-1. DOI: 10.1145/3397481.3450689. URL <https://dl.acm.org/doi/10.1145/3397481.3450689>.
- Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems*, 11(3-4): 24:1–24:45, September 2021b. ISSN 2160-6455. DOI: 10.1145/3387166. URL <https://dl.acm.org/doi/10.1145/3387166>.
- Robert Molloy and Raja Parasuraman. Monitoring an Automated System for a Single Failure: Vigilance and Task Complexity Effects. *Human Factors*, 38(2):311–322, June 1996. ISSN 0018-7208. DOI: 10.1177/001872089606380211. URL <https://doi.org/10.1177/001872089606380211>. Publisher: SAGE Publications Inc.
- Christoph Molnar. *Interpretable Machine Learning*. 2019. URL <https://christophm.github.io/interpretable-ml-book/>.
- David L. Morgan. Focus Groups. *Annual Review of Sociology*, 22(1):129–152, 1996. DOI: 10.1146/annurev.soc.22.1.129. URL <https://doi.org/10.1146/annurev.soc.22.1.129>. _eprint: <https://doi.org/10.1146/annurev.soc.22.1.129>.
- Jessica Morley, Caio C. V. Machado, Christopher Burr, Josh Cowsls, Indra Joshi, Mariarosaria Taddeo, and Luciano Floridi. The ethics of AI in health care: A mapping review. *Social Science & Medicine*, 260:113172, September 2020. ISSN 0277-9536. DOI: 10.1016/j.socscimed.2020.113172. URL <https://www.sciencedirect.com/science/article/pii/S0277953620303919>.
- Cecily Morrison, Kit Huckvale, Bob Corish, Richard Banks, Martin Grayson, Jonas Dorn, Abigail Sellen, and S an Lindley. Visualizing Ubiquitously Sensed Measures of Motor Ability in Multiple Sclerosis: Reflections on Communicating Machine Learning in Practice. *ACM Transactions on Interactive Intelligent Systems*, 8(2):1–28, July 2018. ISSN 2160-6455, 2160-6463. DOI: 10.1145/3181670. URL <https://dl.acm.org/doi/10.1145/3181670>.
- David A. Moss, David Moss, and John Cisternino. *New Perspectives on Regulation*. The Tobin Project, 2009. ISBN 978-0-9824788-0-6. Google-Books-ID: wEQ6QGS6sPkC.
- Stephen Muggleton. Inductive logic programming. *New Generation Computing*, 8(4):295–318, February 1991. ISSN 1882-7055. DOI: 10.1007/BF03037089. URL <https://doi.org/10.1007/BF03037089>.
- C. D. Mulrow. Systematic Reviews: Rationale for systematic reviews. *BMJ*, 309(6954):597–599, September 1994. ISSN 0959-8138, 1468-5833. DOI: 10.1136/bmj.309.6954.597. URL <https://www.bmj.com/content/309/6954/597>. Publisher: British Medical Journal Publishing Group Section: Education and debate.
- Zachary Munn, Micah D. J. Peters, Cindy Stern, Catalin Tufanaru, Alexa McArthur, and Edoardo Aromataris. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology*, 18(1):143, November 2018. ISSN 1471-2288. DOI: 10.1186/s12874-018-0611-x. URL <https://doi.org/10.1186/s12874-018-0611-x>.
- Jakob M okander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. Auditing large language models: a three-layered approach. *AI and Ethics*, 3(2):31, May 2023. ISSN 2730-5961. DOI: 10.1007/s43681-023-00289-2. URL <https://doi.org/10.1007/s43681-023-00289-2>.
- Berndt M uller, Joachim Reinhardt, and Michael T. Strickland. *Neural Networks: An Introduction*. Springer Science & Business Media, October 1995. ISBN 978-3-540-60207-1.
- Mohammad Naiseh, Nan Jiang, Jianbing Ma, and Raian Ali. Personalising Explainable Recommendations: Literature and Conceptualisation. In  lvvaro Rocha, Hojjat Adeli, Lu s Paulo Reis, Sandra Costanzo, Irena Orovic, and Fernando Moreira, editors, *Trends and Innovations in Information Systems and Technologies*, Advances in Intelligent Systems and Computing, pages 518–533, Cham, 2020. Springer International Publishing. ISBN 978-3-030-45691-7.

- Mohammad Naiseh, Reem S. Al-Mansoori, Dena Al-Thani, Nan Jiang, and Raian Ali. Nudging through Friction: An Approach for Calibrating Trust in Explainable AI. In *2021 8th International Conference on Behavioral and Social Computing (BESC)*, pages 1–5, October 2021a. DOI: 10.1109/BESC53957.2021.9635271.
- Mohammad Naiseh, Deniz Cemiloglu, Dena Al Thani, Nan Jiang, and Raian Ali. Explainable Recommendations and Calibrated Trust: Two Systematic User Errors. *Computer*, 54(10):28–37, October 2021b. ISSN 1558-0814. DOI: 10.1109/MC.2021.3076131. Conference Name: Computer.
- Luca Nannini, Agathe Balayn, and Adam Leon Smith. Explainability in AI Policies: A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23*, pages 1198–1212, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. DOI: 10.1145/3593013.3594074. URL <https://dl.acm.org/doi/10.1145/3593013.3594074>.
- National Transportation Safety Board. Aircraft Accident Report, In-flight Breakup Over the Atlantic Ocean, Trans World Airlines Flight 800, Boeing 747-131, N93119 Near East Moriches, New York, July 17, 1996. Technical report, August 2000.
- Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlöterer, Maurice van Keulen, and Christin Seifert. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Computing Surveys*, 55(13s):295:1–295:42, 2023. ISSN 0360-0300. DOI: 10.1145/3583558. URL <https://dl.acm.org/doi/10.1145/3583558>.
- E. W. T. Ngai, Yong Hu, Y. H. Wong, Yijun Chen, and Xin Sun. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3):559–569, February 2011. ISSN 0167-9236. DOI: 10.1016/j.dss.2010.08.006. URL <https://www.sciencedirect.com/science/article/pii/S0167923610001302>.
- Anh Nguyen, Jason Yosinski, and Jeff Clune. Multifaceted Feature Visualization: Uncovering the Different Types of Features Learned By Each Neuron in Deep Neural Networks, May 2016. URL <http://arxiv.org/abs/1602.03616>. arXiv:1602.03616 [cs].
- Jakob Nielsen. Finding usability problems through heuristic evaluation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '92*, pages 373–380, New York, NY, USA, 1992. Association for Computing Machinery. ISBN 978-0-89791-513-7. DOI: 10.1145/142750.142834. URL <https://dl.acm.org/doi/10.1145/142750.142834>.
- NIST. AI Risk Management Framework: AI RMF (1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology, Gaithersburg, MD, 2023. URL <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>.
- Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *26th International Conference on Intelligent User Interfaces*, pages 340–350, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8017-1. URL <https://doi.org/10.1145/3397481.3450639>.
- Brendan Nyhan and Jason Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, 2010. ISSN 1573-6687. DOI: 10.1007/s11109-010-9112-2. Place: Germany Publisher: Springer.
- Jonathan A. Obar. Sunlight alone is not a disinfectant: Consent and the futility of opening Big Data black boxes (without assistance). *Big Data & Society*, 7(1):2053951720935615, January 2020. ISSN 2053-9517. DOI: 10.1177/2053951720935615. URL <https://doi.org/10.1177/2053951720935615>. Publisher: SAGE Publications Ltd.
- Heather L. O'Brien and Elaine G. Toms. What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, 59(6):938–955, 2008. ISSN 1532-2890. DOI: 10.1002/asi.20801. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20801>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.20801>.

- OECD. Recommendation of the Council on Artificial Intelligence. Technical report, OECD, May 2019. URL <https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449>.
- OECD. Transparency and explainability (OECD AI Principle) - OECD.AI, 2019. URL <https://oecd.ai/en/dashboards/ai-principles/P7>.
- OECD. *OECD Business and Finance Outlook 2021: AI in Business and Finance, Chapter 5: The use of SupTech to enhance market supervision and integrity*. OECD Business and Finance Outlook. OECD, September 2021a. ISBN 978-92-64-64469-4 978-92-64-70629-3 978-92-64-57363-5 978-92-64-76483-5. DOI: 10.1787/ba682899-en. URL https://www.oecd-ilibrary.org/finance-and-investment/oecd-business-and-finance-outlook-2021_ba682899-en.
- OECD. Risk-based regulation. In *OECD Regulatory Policy Outlook 2021*. OECD, October 2021b. ISBN 978-92-64-94868-6 978-92-64-80247-6 978-92-64-87415-2 978-92-64-52892-5. DOI: 10.1787/9d082a11-en. URL https://www.oecd-ilibrary.org/governance/oecd-regulatory-policy-outlook-2021_9d082a11-en.
- Jeroen Ooge. *Explaining Artificial Intelligence With Tailored Interactive Visualisations*. PhD thesis, October 2023.
- Jeroen Ooge, Shotallo Kato, and Katrien Verbert. Explaining Recommendations in E-Learning: Effects on Adolescents' Trust. In *27th International Conference on Intelligent User Interfaces, IUI '22*, pages 93–105, New York, NY, USA, March 2022. Association for Computing Machinery. ISBN 978-1-4503-9144-3. DOI: 10.1145/3490099.3511140. URL <https://doi.org/10.1145/3490099.3511140>.
- Antti Oulasvirta, Jussi P. P. Jokinen, and Andrew Howes. Computational Rationality as a Theory of Interaction. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22*, pages 1–14, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 978-1-4503-9157-3. DOI: 10.1145/3491102.3517739. URL <https://dl.acm.org/doi/10.1145/3491102.3517739>.
- Erik Overrein. How machine learning can dramatically reduce financial institutions' cost of compliance, May 2020. URL <https://www.bearingpoint.com/en-no/insights-events/insights/machine-learning-is-the-key-to-efficient-and-effective-aml/>. Accessed 8/27/2023.
- Heather O'Brien and Paul Cairns. An empirical evaluation of the User Engagement Scale (UES) in online news environments. *Information Processing & Management*, 51(4):413–427, July 2015. ISSN 0306-4573. DOI: 10.1016/j.ipm.2015.03.003. URL <https://www.sciencedirect.com/science/article/pii/S0306457315000412>.
- Heather L. O'Brien, Paul Cairns, and Mark Hall. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies*, 112:28–39, April 2018. ISSN 1071-5819. DOI: 10.1016/j.ijhcs.2018.01.004. URL <https://www.sciencedirect.com/science/article/pii/S1071581918300041>.
- Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, and others. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Systematic reviews*, 10(1):1–11, 2021. Publisher: BioMed Central.
- Cecilia Panigutti, Andrea Beretta, Daniele Fadda, Fosca Giannotti, Dino Pedreschi, Alan Perotti, and Salvatore Rinzivillo. Co-design of Human-centered, Explainable AI for Clinical Decision Support. *ACM Transactions on Interactive Intelligent Systems*, 13(4):21:1–21:35, 2023a. ISSN 2160-6455. DOI: 10.1145/3587271. URL <https://dl.acm.org/doi/10.1145/3587271>.
- Cecilia Panigutti, Ronan Hamon, Isabelle Hupont, David Fernandez Llorca, Delia Fano Yela, Henrik Junklewitz, Salvatore Scalzo, Gabriele Mazzini, Ignacio Sanchez, Josep Soler Garrido, and Emilia Gomez. The role of explainable AI in the context of the AI Act. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1139–1150, Chicago IL USA, June 2023b. ACM. ISBN 9798400701924. DOI: 10.1145/3593013.3594069. URL <https://dl.acm.org/doi/10.1145/3593013.3594069>.

- Raja Parasuraman and Dietrich H. Manzey. Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors*, 52(3):381–410, June 2010. ISSN 0018-7208. DOI: 10.1177/0018720810376055. URL <https://doi.org/10.1177/0018720810376055>. Publisher: SAGE Publications Inc.
- Raja Parasuraman and Victor Riley. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors*, 39(2):230–253, June 1997. ISSN 0018-7208. URL <https://doi.org/10.1518/001872097778543886>. Publisher: SAGE Publications Inc.
- Raja Parasuraman, Robert Molloy, and Indramani Singh. Performance Consequences of Automation Induced Complacency. *International Journal of Aviation Psychology*, 3, February 1993.
- Hyanghee Park, Daehwan Ahn, Kartik Hosanagar, and Joonhwan Lee. Human-AI Interaction in Human Resource Management: Understanding Why Employees Resist Algorithmic Evaluation at Workplaces and How to Mitigate Burdens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–15, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8096-6. DOI: 10.1145/3411764.3445304. URL <https://doi.org/10.1145/3411764.3445304>.
- Frank Pasquale. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press, 2015. ISBN 978-0-674-36827-9. URL <https://www.jstor.org/stable/j.ctt13x0hch>.
- Paul Fehlinger. Enabling the responsible use of technology at scale. Technical report, SITRA, October 2023. URL <https://www.sitra.fi/en/publications/enabling-the-responsible-use-of-technology-at-scale/>.
- Paul Thagard. Explanatory coherence. *Behavioral and brain sciences*, 12:435–502, 1989.
- Georgios Pavlidis. Deploying artificial intelligence for anti-money laundering and asset recovery: the dawn of a new era. *Journal of Money Laundering Control*, 26(7):155–166, January 2023. ISSN 1368-5201. DOI: 10.1108/JMLC-03-2023-0050. URL <https://doi.org/10.1108/JMLC-03-2023-0050>. Publisher: Emerald Publishing Limited.
- Sean Penney, Jonathan Dodge, Claudia Hilderbrand, Andrew Anderson, Logan Simpson, and Margaret Burnett. Toward Foraging for Understanding of StarCraft Agents: An Empirical Study. In *23rd International Conference on Intelligent User Interfaces*, IUI '18, pages 225–237, New York, NY, USA, March 2018. Association for Computing Machinery. ISBN 978-1-4503-4945-1. DOI: 10.1145/3172944.3172946. URL <https://doi.org/10.1145/3172944.3172946>.
- Nancy Pennington and Reid Hastie. Reasoning in explanation-based decision making. *Cognition*, 49(1): 123–163, October 1993. ISSN 0010-0277. DOI: 10.1016/0010-0277(93)90038-W. URL <https://www.sciencedirect.com/science/article/pii/001002779390038W>.
- Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, October 2017. ISBN 978-0-262-03731-0.
- Lawrence D. Phillips and Ward Edwards. Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, 72(3):346–354, 1966. ISSN 0022-1015. DOI: 10.1037/h0023653. Place: US Publisher: American Psychological Association.
- Sayantan Polley, Suhita Ghosh, Marcus Thiel, Michael Kotzyba, and Andreas Nürnberger. SIMFIC: An Explainable Book Search Companion. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, pages 1–6, September 2020. DOI: 10.1109/ICHMS49158.2020.9209581.
- David Poole, Randy Goebel, and Romas Aleliunas. Theorist: A Logical Reasoning System for Defaults and Diagnosis. In *The Knowledge Frontier*, pages 331–352. Springer New York, New York, NY, 1987.
- Harry E. Pople. On the mechanization of abductive logic. In *Proceedings of the 3rd international joint conference on Artificial intelligence*, IJCAI'73, pages 147–152, San Francisco, CA, USA, 1973. Morgan Kaufmann Publishers Inc.

- Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and Measuring Model Interpretability. *arXiv:1802.07810 [cs]*, November 2019. URL <http://arxiv.org/abs/1802.07810>. arXiv: 1802.07810.
- Forough Poursabzi-Sangdeh, Samira Samadi, Jennifer Wortman Vaughan, and Hanna Wallach. A Human in the Loop is Not Enough: The Need for Human-Subject Experiments in Facial Recognition. May 2020. URL <https://www.microsoft.com/en-us/research/publication/a-human-in-the-loop-is-not-enough-the-need-for-human-subject-experiments-in-facial-recognition/>.
- Aimee Prawitz, E. Thomas Garman, Benoit Sorhaindo, Barbara O'Neill, Jinhee Kim, and Patricia Drentea. Incharge Financial Distress/Financial Well-Being Scale: Development, Administration, and Score Interpretation, 2006. URL <https://papers.ssrn.com/abstract=2239338>.
- Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating Training Data Influence by Tracing Gradient Descent. In *Advances in Neural Information Processing Systems*, volume 33, pages 19920–19930. Curran Associates, Inc., 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/e6385d39ec9394f2f3a354d9d2b88eec-Abstract.html>.
- Inioluwa Deborah Raji and Joy Buolamwini. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 429–435, Honolulu HI USA, January 2019. ACM. ISBN 978-1-4503-6324-2. DOI: 10.1145/3306618.3314244. URL <https://dl.acm.org/doi/10.1145/3306618.3314244>.
- Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 33–44, Barcelona Spain, January 2020. ACM. ISBN 978-1-4503-6936-7. DOI: 10.1145/3351095.3372873. URL <https://dl.acm.org/doi/10.1145/3351095.3372873>.
- Gabriëlle Ras, Marcel van Gerven, and Pim Haselager. Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges. In Hugo Jair Escalante, Sergio Escalera, Isabelle Guyon, Xavier Baró, Yağmur Güçlütürk, Umut Güçlü, and Marcel van Gerven, editors, *Explainable and Interpretable Models in Computer Vision and Machine Learning*, The Springer Series on Challenges in Machine Learning, pages 19–36. Springer International Publishing, Cham, 2018. ISBN 978-3-319-98131-4. URL https://doi.org/10.1007/978-3-319-98131-4_2.
- Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R. Varshney, Amit Dhurandhar, and Richard Tomsett. Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making. *arXiv:2010.07938 [cs]*, October 2020. URL <http://arxiv.org/abs/2010.07938>. arXiv: 2010.07938.
- Stephen J. Read and Amy Marcus-Newhall. Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, 65(3):429–447, 1993. ISSN 1939-1315. DOI: 10.1037/0022-3514.65.3.429. Place: US Publisher: American Psychological Association.
- Juan Rebanal, Jordan Combitsis, Yuqi Tang, and Xiang ‘Anthony’ Chen. XAlgo: a Design Probe of Explaining Algorithms’ Internal States via Question-Answering. In *26th International Conference on Intelligent User Interfaces, IUI ’21*, pages 329–339, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8017-1. DOI: 10.1145/3397481.3450676. URL <https://doi.org/10.1145/3397481.3450676>.
- Chris Reed. How should we regulate artificial intelligence? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128):20170360, August 2018. DOI: 10.1098/rsta.2017.0360. URL <https://royalsocietypublishing.org/doi/10.1098/rsta.2017.0360>. Publisher: Royal Society.
- Bob Rehder. When similarity and causality compete in category-based property generalization. *Memory & Cognition*, 34(1):3–16, January 2006. ISSN 0090-502X. DOI: 10.3758/bf03193382.

- Dent M. Rhodes and Janet White Azbell. Designing Interactive Video Instruction Professionally. *Training and Development Journal*, 39(12):31–33, 1985.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "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*, KDD '16, pages 1135–1144, New York, NY, USA, August 2016. Association for Computing Machinery. ISBN 978-1-4503-4232-2. DOI: 10.1145/2939672.2939778. URL <https://doi.org/10.1145/2939672.2939778>.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. ISSN 2374-3468. URL <https://ojs.aaai.org/index.php/AAAI/article/view/11491>. Number: 1.
- Mireia Ribera and Agata Lapedriza. Can we do better explanations? A proposal of User-Centered Explainable AI. *Los Angeles*, page 7, 2019.
- Yvonne Rogers, Helen Sharp, and Jenny Preece. *Interaction Design: beyond human-computer interaction (6th edition)*. John Wiley & Sons, March 2023. ISBN 978-1-119-90109-9. URL <https://oro.open.ac.uk/88758/>.
- Katharina J. Rohlfing, Philipp Cimiano, Ingrid Scharlau, Tobias Matzner, Heike M. Buhl, Hendrik Buschmeier, Elena Esposito, Angela Grimminger, Barbara Hammer, Reinhold Häb-Umbach, Ilona Horwath, Eyke Hüllermeier, Friederike Kern, Stefan Kopp, Kirsten Thommes, Axel-Cyrille Ngonga Ngomo, Carsten Schulte, Henning Wachsmuth, Petra Wagner, and Britta Wrede. Explanation as a Social Practice: Toward a Conceptual Framework for the Social Design of AI Systems. *IEEE Transactions on Cognitive and Developmental Systems*, 13(3):717–728, September 2021. ISSN 2379-8939. DOI: 10.1109/TCDS.2020.3044366. URL <https://ieeexplore.ieee.org/document/9292993>. Conference Name: IEEE Transactions on Cognitive and Developmental Systems.
- Avi Rosenfeld and Ariella Richardson. Explainability in human-agent systems. *Autonomous Agents and Multi-Agent Systems*, 33(6):673–705, November 2019. ISSN 1573-7454. DOI: 10.1007/s10458-019-09408-y. URL <https://doi.org/10.1007/s10458-019-09408-y>.
- Andrew Ross, Nina Chen, Elisa Zhao Hang, Elena L. Glassman, and Finale Doshi-Velez. Evaluating the Interpretability of Generative Models by Interactive Reconstruction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, Yokohama Japan, May 2021. ACM. ISBN 978-1-4503-8096-6. DOI: 10.1145/3411764.3445296. URL <https://dl.acm.org/doi/10.1145/3411764.3445296>.
- Mary Beth Rosson and John M. Carroll. Scenario-based design. In *Human-computer Interaction*, page 20. CRC Press, Boca Raton, 1st edition edition, March 2009. ISBN 978-0-429-13939-0. DOI: 10.1201/9781420088892-14. URL <https://www.taylorfrancis.com/chapters/edit/10.1201/9781420088892-14/scenario-based-design-mary-beth-rosson-john-carroll>. Pages: 161-180 Publication Title: Human-Computer Interaction.
- Paul A. Roth. How Narratives Explain. *Social Research*, 56(2):449–478, 1989. ISSN 0037-783X. URL <https://www.jstor.org/stable/40970551>. Publisher: The New School.
- Steven F. Roth and Joe Mattis. Data characterization for intelligent graphics presentation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '90, pages 193–200, New York, NY, USA, March 1990. Association for Computing Machinery. ISBN 978-0-201-50932-8. DOI: 10.1145/97243.97273. URL <https://doi.org/10.1145/97243.97273>.
- Denise M. Rousseau, Sim B. Sitkin, Ronald S. Burt, and Colin Camerer. Introduction to Special Topic Forum: Not so Different after All: A Cross-Discipline View of Trust. *The Academy of Management Review*, 23(3):393–404, 1998. URL <http://www.jstor.org/stable/259285>.
- Maria Roussou. Learning by doing and learning through play: an exploration of interactivity in virtual environments for children. *Computers in Entertainment*, 2(1):10, January 2004. DOI: 10.1145/973801.973818. URL <https://doi.org/10.1145/973801.973818>.

- Antoinette Rouvroy. The end(s) of critique: Data behaviourism versus due process. In *Privacy Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology*, pages 143–167. Taylor & Francis, 2013. ISBN 978-0-203-42764-4. DOI: 10.4324/9780203427644. URL <http://www.scopus.com/inward/record.url?scp=84917399654&partnerID=8YFLogxK>.
- Hofit Wasserman Rozen, Niva Elkin-Koren, and Ran Gilad-Bachrach. The Case Against Explainability, May 2023. URL <http://arxiv.org/abs/2305.12167>. arXiv:2305.12167 [cs].
- Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019. ISSN 2522-5839. DOI: 10.1038/s42256-019-0048-x. URL <https://www.nature.com/articles/s42256-019-0048-x>. Number: 5 Publisher: Nature Publishing Group.
- Stuart J Russell and Peter Norvig. *Artificial intelligence a modern approach*. 2010.
- Christian Sandvig, Kevin Hamilton, K. Karahalios, and Cédric Langbort. Auditing Algorithms : Research Methods for Detecting Discrimination on Internet Platforms. In *Preconference at the 64th Annual Meeting of the International Communication Association*, page 23, Seattle, WA, USA, May 2014. University of Michigan.
- James Schaffer, Prasanna Giridhar, Debra Jones, Tobias Höllerer, Tarek Abdelzaher, and John O’Donovan. Getting the Message? A Study of Explanation Interfaces for Microblog Data Analysis. In *Proceedings of the 20th International Conference on Intelligent User Interfaces, IUI ’15*, pages 345–356, New York, NY, USA, March 2015. Association for Computing Machinery. ISBN 978-1-4503-3306-1. DOI: 10.1145/2678025.2701406. URL <https://doi.org/10.1145/2678025.2701406>.
- James Schaffer, John O’Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. I can do better than your AI: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI ’19*, pages 240–251, New York, NY, USA, March 2019. Association for Computing Machinery. ISBN 978-1-4503-6272-6. DOI: 10.1145/3301275.3302308. URL <https://doi.org/10.1145/3301275.3302308>.
- Frederick Schauer. Giving Reasons. *Stanford Law Review*, 47(4):633–659, 1995. ISSN 0038-9765. DOI: 10.2307/1229080. URL <https://www.jstor.org/stable/1229080>. Publisher: Stanford Law Review.
- Frederick Schauer. Transparency in Three Dimensions. *University of Illinois Law Review*, 2011:1339, 2011. URL <https://heinonline.org/HOL/Page?handle=hein.journals/unilllr2011&id=1347&div=&collection=>.
- Kara Schick-Makaroff, Marjorie MacDonald, Marilyn Plummer, Judy Burgess, and Wendy Neander. What Synthesis Methodology Should I Use? A Review and Analysis of Approaches to Research Synthesis. *AIMS public health*, 3(1):172–215, March 2016. ISSN 2327-8994. DOI: 10.3934/publichealth.2016.1.172. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5690272/>.
- Johanes Schneider and Joshua Handali. Personalized explanation in machine learning: A conceptualization. In *Proceedings of the European Conference on Information Systems, ECIS 2019*, Stockholm-Uppsala, Sweden, June 2019. arXiv. DOI: 10.48550/arXiv.1901.00770. URL https://aisel.aisnet.org/ecis2019_rp/171. arXiv:1901.00770 [cs, stat].
- Jonas Schuett. Defining the scope of AI regulations. *Law, Innovation and Technology*, 15(1):60–82, January 2023. ISSN 1757-9961. DOI: 10.1080/17579961.2023.2184135. URL <https://doi.org/10.1080/17579961.2023.2184135>. Publisher: Routledge _eprint: <https://doi.org/10.1080/17579961.2023.2184135>.
- Richard Schwier and Earl R. Misanchuk. *Interactive Multimedia Instruction*. Educational Technology, 1993. ISBN 978-0-87778-251-3.
- Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. What makes a good conversation? How controllable attributes affect human judgments, April 2019. URL <http://arxiv.org/abs/1902.08654>. arXiv:1902.08654 [cs].

- Andrew D. Selbst and Solon Barocas. The Intuitive Appeal of Explainable Machines, March 2018. URL <https://papers.ssrn.com/abstract=3126971>.
- Rita Sevastjanova, Wolfgang Jentner, Fabian Sperrle, Rebecca Kehlbeck, Jürgen Bernard, and Menatallah El-assady. QuestionComb: A Gamification Approach for the Visual Explanation of Linguistic Phenomena through Interactive Labeling. *ACM Transactions on Interactive Intelligent Systems*, 11(3-4):1–38, December 2021. ISSN 2160-6455, 2160-6463. DOI: 10.1145/3429448. URL <https://dl.acm.org/doi/10.1145/3429448>.
- Patrick Shafto and John D. Coley. Development of categorization and reasoning in the natural world: novices to experts, naive similarity to ecological knowledge. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 29(4):641–649, July 2003. ISSN 0278-7393. DOI: 10.1037/0278-7393.29.4.641.
- Lei Shi, Zhiyang Teng, Le Wang, Yue Zhang, and Alexander Binder. DeepClue: Visual Interpretation of Text-Based Deep Stock Prediction. *IEEE Transactions on Knowledge and Data Engineering*, 31(6):1094–1108, June 2019. ISSN 1558-2191. DOI: 10.1109/TKDE.2018.2854193. Conference Name: IEEE Transactions on Knowledge and Data Engineering.
- Asaya Shimojo, Kazuhisa Miwa, and Hitoshi Terai. How Does Explanatory Virtue Determine Probability Estimation?—Empirical Discussion on Effect of Instruction. *Frontiers in Psychology*, 11, 2020. ISSN 1664-1078. URL <https://www.frontiersin.org/article/10.3389/fpsyg.2020.575746>.
- Dajung Diane Shin and Sung-il Kim. Homo Curious: Curious or Interested? *Educational Psychology Review*, 31(4):853–874, December 2019. ISSN 1573-336X. DOI: 10.1007/s10648-019-09497-x. URL <https://doi.org/10.1007/s10648-019-09497-x>.
- Donghee Shin. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, 146:102551, February 2021. ISSN 1071-5819. DOI: 10.1016/j.ijhcs.2020.102551. URL <https://www.sciencedirect.com/science/article/pii/S1071581920301531>.
- Ben Shneiderman. Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. *ACM Transactions on Interactive Intelligent Systems*, 10(4):26:1–26:31, October 2020. ISSN 2160-6455. DOI: 10.1145/3419764. URL <https://dl.acm.org/doi/10.1145/3419764>.
- Auste Simkute, Ewa Luger, Mike Evans, and Rhianne Jones. Experts in the Shadow of Algorithmic Systems: Exploring Intelligibility in a Decision-Making Context. In *Companion Publication of the 2020 ACM Designing Interactive Systems Conference, DIS' 20 Companion*, pages 263–268, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 978-1-4503-7987-8. DOI: 10.1145/3393914.3395862. URL <https://doi.org/10.1145/3393914.3395862>.
- Daniel J. Simons. Current Approaches to Change Blindness. *Visual Cognition*, 7(1-3):1–15, January 2000. ISSN 1350-6285. DOI: 10.1080/135062800394658. URL <https://doi.org/10.1080/135062800394658>. Publisher: Routledge_eprint: <https://doi.org/10.1080/135062800394658>.
- Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, April 2014. URL <http://arxiv.org/abs/1312.6034>. arXiv:1312.6034 [cs].
- Rod Sims. Interactivity: A forgotten art? *Computers in Human Behavior*, 13(2):157–180, May 1997. ISSN 0747-5632. DOI: 10.1016/S0747-5632(97)00004-6. URL <https://www.sciencedirect.com/science/article/pii/S0747563297000046>.
- Radish Singh, Miguel Fernandes, Nick Lim, and Eric Ang. The case for artificial intelligence in combating money laundering and terrorist financing. Technical report, Deloitte, 2018. URL <https://www2.deloitte.com/mm/en/pages/financial-advisory/articles/the-case-for-artificial-intelligence-in-combating-money-laundering-and-terrorist-financing.html>.

- Frédérique Six. Trust in Regulatory Relations. *Public Management Review*, 15(2):163–185, February 2013. ISSN 1471-9037. DOI: 10.1080/14719037.2012.727461. URL <https://doi.org/10.1080/14719037.2012.727461>. Publisher: Routledge_eprint: <https://doi.org/10.1080/14719037.2012.727461>.
- Dylan Slack, Anna Hilgard, Sameer Singh, and Himabindu Lakkaraju. Reliable Post hoc Explanations: Modeling Uncertainty in Explainability. In *Advances in Neural Information Processing Systems*, volume 34, pages 9391–9404. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/4e246a381baf2ce038b3b0f82c7d6fb4-Abstract.html.
- Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. TalkToModel: Explaining Machine Learning Models with Interactive Natural Language Conversations, September 2022. URL <http://arxiv.org/abs/2207.04154>. arXiv:2207.04154 [cs].
- Nathalie A. Smuha. From a ‘race to AI’ to a ‘race to AI regulation’: regulatory competition for artificial intelligence. *Law, Innovation and Technology*, 13(1):57–84, January 2021. ISSN 1757-9961. DOI: 10.1080/17579961.2021.1898300. URL <https://doi.org/10.1080/17579961.2021.1898300>. Publisher: Routledge_eprint: <https://doi.org/10.1080/17579961.2021.1898300>.
- Dominic S.B. Soh and Nonna Martinov-Bennie. The internal audit function: Perceptions of internal audit roles, effectiveness and evaluation. *Managerial Auditing Journal*, 26(7):605–622, January 2011. ISSN 0268-6902. DOI: 10.1108/02686901111151332. URL <https://doi.org/10.1108/02686901111151332>. Publisher: Emerald Group Publishing Limited.
- Kacper Sokol and Peter Flach. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 56–67, Barcelona Spain, January 2020. ACM. ISBN 978-1-4503-6936-7. DOI: 10.1145/3351095.3372870. URL <http://dl.acm.org/doi/10.1145/3351095.3372870>.
- Francesco Sovrano and Fabio Vitali. From Philosophy to Interfaces: an Explanatory Method and a Tool Inspired by Achinstein’s Theory of Explanation. In *26th International Conference on Intelligent User Interfaces*, pages 81–91, College Station TX USA, April 2021. ACM. ISBN 978-1-4503-8017-1. DOI: 10.1145/3397481.3450655. URL <https://dl.acm.org/doi/10.1145/3397481.3450655>.
- Thilo Spinner, Udo Schlegel, Hanna Schäfer, and Mennatallah El-Assady. explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1064–1074, January 2020. ISSN 1941-0506. DOI: 10.1109/TVCG.2019.2934629. Conference Name: IEEE Transactions on Visualization and Computer Graphics.
- Clay Spinuzzi. The Methodology of Participatory Design. *Technical Communication*, 52(2):163–174, May 2005.
- Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for Simplicity: The All Convolutional Net, April 2015. URL <http://arxiv.org/abs/1412.6806>. arXiv:1412.6806 [cs].
- Aaron Springer and Steve Whittaker. Progressive disclosure: empirically motivated approaches to designing effective transparency. In *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI ’19*, pages 107–120, New York, NY, USA, March 2019. Association for Computing Machinery. ISBN 978-1-4503-6272-6. DOI: 10.1145/3301275.3302322. URL <https://doi.org/10.1145/3301275.3302322>.
- Brian Stanton and Theodore Jensen. Trust and Artificial Intelligence. preprint, March 2021. URL <https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8332-draft.pdf>.
- Constantine Stephanidis, Gavriel Salvendy, Demosthenes Akoumianakis, Albert Arnold, Nigel Bevan, Daniel Dardailler, Pier Luigi Emiliani, Ilias Iakovidis, Phil Jenkins, Arthur Karshmer, Peter Korn, Aaron Marcus, Harry Murphy, Charles Oppermann, Christian Stary, Hiroshi Tamura, Manfred Tscheligi, Hirotada Ueda, Gerhard Weber, and Juergen Ziegler. Toward an Information Society for All: HCI Challenges and R&D Recommendations. *International Journal of Human–Computer Interaction*, 11(1): 1–28, January 1999. ISSN 1044-7318. URL https://doi.org/10.1207/s15327590ijhc1101_1.

- Ilija Stepin, Jose M. Alonso, Alejandro Catala, and Martín Pereira-Fariña. A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. *IEEE Access*, 9: 11974–12001, 2021. ISSN 2169-3536. DOI: 10.1109/ACCESS.2021.3051315. URL <https://ieeexplore.ieee.org/document/9321372>. Conference Name: IEEE Access.
- Jonathan Steuer. Defining Virtual Reality: Dimensions Determining Telepresence. *Journal of Communication*, pages 73–93, 1992.
- Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies*, 67(8):639–662, August 2009. ISSN 1071-5819. DOI: 10.1016/j.ijhcs.2009.03.004. URL <https://www.sciencedirect.com/science/article/pii/S1071581909000457>.
- Mark C. Suchman. Managing Legitimacy: Strategic and Institutional Approaches. *The Academy of Management Review*, 20(3):571–610, 1995. ISSN 0363-7425. DOI: 10.2307/258788. URL <https://www.jstor.org/stable/258788>. Publisher: Academy of Management.
- Jiao Sun, Q. Vera Liao, Michael Muller, Mayank Agarwal, Stephanie Houde, Kartik Talamadupula, and Justin D. Weisz. Investigating Explainability of Generative AI for Code through Scenario-based Design. In *27th International Conference on Intelligent User Interfaces, IUI '22*, pages 212–228, New York, NY, USA, March 2022. Association for Computing Machinery. ISBN 978-1-4503-9144-3. DOI: 10.1145/3490099.3511119. URL <https://doi.org/10.1145/3490099.3511119>.
- S. Shyam Sundar, Qian Xu, and Saraswathi Bellur. Designing interactivity in media interfaces: a communications perspective. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '10*, pages 2247–2256, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 978-1-60558-929-9. DOI: 10.1145/1753326.1753666. URL <https://doi.org/10.1145/1753326.1753666>.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks, June 2017. URL <http://arxiv.org/abs/1703.01365>. arXiv:1703.01365 [cs].
- Harry Surden. Artificial Intelligence and Law: An Overview, June 2019. URL <https://papers.ssrn.com/abstract=3411869>.
- Harini Suresh, Steven R. Gomez, Kevin K. Nam, and Arvind Satyanarayan. Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21*, pages 1–16, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8096-6. DOI: 10.1145/3411764.3445088. URL <https://doi.org/10.1145/3411764.3445088>.
- Harini Suresh, Kathleen M Lewis, John Guttag, and Arvind Satyanarayan. Intuitively Assessing ML Model Reliability through Example-Based Explanations and Editing Model Inputs. In *27th International Conference on Intelligent User Interfaces, IUI '22*, pages 767–781, New York, NY, USA, March 2022. Association for Computing Machinery. ISBN 978-1-4503-9144-3. DOI: 10.1145/3490099.3511160. URL <https://doi.org/10.1145/3490099.3511160>.
- Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. Visual, textual or hybrid: the effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces*, pages 109–119, College Station TX USA, April 2021. ACM. ISBN 978-1-4503-8017-1. DOI: 10.1145/3397481.3450662. URL <https://dl.acm.org/doi/10.1145/3397481.3450662>.
- Stefano Teso, Öznur Alkan, Wolfgang Stammer, and Elizabeth Daly. Leveraging explanations in interactive machine learning: An overview. *Frontiers in Artificial Intelligence*, 6, 2023. ISSN 2624-8212. URL <https://www.frontiersin.org/articles/10.3389/frai.2023.1066049>.
- Paul Thagard. Explanatory coherence. *Behavioral and Brain Sciences*, 12(3):435–467, September 1989. ISSN 1469-1825, 0140-525X. DOI: 10.1017/S0140525X00057046. URL <https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/abs/explanatory-coherence/E05CB61CD64C26138E794BC601CC9D7A>. Publisher: Cambridge University Press.

- The Federal Reserve Board of Governors in Washington DC. The Fed - Supervisory Letter SR 11-7 on guidance on Model Risk Management, April 2011. URL <https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm>.
- The White House. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. October 2023. URL <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/>.
- Arthur Thuy and Dries F. Benoit. Explainability through uncertainty: Trustworthy decision-making with neural networks. *European Journal of Operational Research*, September 2023. ISSN 0377-2217. DOI: 10.1016/j.ejor.2023.09.009. URL <https://www.sciencedirect.com/science/article/pii/S0377221723007105>.
- Nava Tintarev. Explanations of recommendations. In *Proceedings of the 2007 ACM conference on Recommender systems*, RecSys '07, pages 203–206, New York, NY, USA, October 2007. Association for Computing Machinery. ISBN 978-1-59593-730-8. DOI: 10.1145/1297231.1297275. URL <https://doi.org/10.1145/1297231.1297275>.
- Adeline Toader. Auditability of AI Systems – Brake or Acceleration to Innovation?, November 2019. URL <https://papers.ssrn.com/abstract=3526222>.
- Dylan Tokar. Google Cloud Launches Anti-Money-Laundering Tool for Banks, Betting on the Power of AI. *Wall Street Journal*, June 2023. ISSN 0099-9660. URL <https://www.wsj.com/articles/google-cloud-launches-anti-money-laundering-tool-for-banks-betting-on-the-power-of-ai-2512ccce>.
- Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems. In *2018 ICML Workshop on Human Interpretability in Machine Learning*, page 7, Stockholm, Sweden, June 2018. arXiv. URL <http://arxiv.org/abs/1806.07552>. arXiv: 1806.07552.
- Trade and Industry Appeals Tribunal. *Bunq vs. DNB*, ECLI:NL:CBB:2022:707, 21/323 and 21/1108, October 2022. URL <https://deemlink.rechtspraak.nl/uitspraak?id=ECLI:NL:CBB:2022:707>. Soort: Uitspraak.
- Andrea C. Tricco, Erin Lillie, Wasifa Zarin, Kelly K. O'Brien, Heather Colquhoun, Danielle Levac, David Moher, Micah D.J. Peters, Tanya Horsley, Laura Weeks, Susanne Hempel, Elie A. Akl, Christine Chang, Jessie McGowan, Lesley Stewart, Lisa Hartling, Adrian Aldcroft, Michael G. Wilson, Chantelle Garritty, Simon Lewin, Christina M. Godfrey, Marilyn T. Macdonald, Etienne V. Langlois, Karla Soares-Weiser, Jo Moriarty, Tammy Clifford, Özge Tunçalp, and Sharon E. Straus. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Annals of Internal Medicine*, 169(7):467–473, October 2018. ISSN 0003-4819. DOI: 10.7326/M18-0850. URL <https://www.acpjournals.org/doi/10.7326/M18-0850>. Publisher: American College of Physicians.
- Jon Truby, Rafael Brown, and Andrew Dahdal. Banking on AI: mandating a proactive approach to AI regulation in the financial sector. *Law and Financial Markets Review*, 14(2):110–120, April 2020. ISSN 1752-1440. DOI: 10.1080/17521440.2020.1760454. URL <https://doi.org/10.1080/17521440.2020.1760454>. Publisher: Routledge _eprint: <https://doi.org/10.1080/17521440.2020.1760454>.
- Chun-Hua Tsai, Yue You, Xinning Gui, Yubo Kou, and John M. Carroll. Exploring and Promoting Diagnostic Transparency and Explainability in Online Symptom Checkers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–17, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8096-6. DOI: 10.1145/3411764.3445101. URL <https://doi.org/10.1145/3411764.3445101>.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting, December 2023. URL <http://arxiv.org/abs/2305.04388>. arXiv:2305.04388 [cs].

- Alec Tyson and Emma Kikuchi. Growing public concern about the role of artificial intelligence in daily life, August 2023. URL <https://www.pewresearch.org/short-reads/2023/08/28/growing-public-concern-about-the-role-of-artificial-intelligence-in-daily-life/>.
- Brigitte Unger and Elena Madalina Busuioc. *The Scale and Impacts of Money Laundering*. Edward Elgar Publishing, March 2007. ISBN 978-1-78100-762-4.
- UNODC. Estimating illicit financial flows resulting from drug trafficking and other transnational organized crimes. Discussion paper, United Nations, October 2011. URL https://www.unodc.org/documents/data-and-analysis/Studies/Illicit_financial_flows_2011_web.pdf.
- Amy Unruh and Sarah Robinson. Explaining Model Predictions On Structured Data, March 2020. URL <https://liwaiwai.com/2020/03/04/explaining-model-predictions-on-structured-data/>.
- Betty Vandenbosch and Michael J. Ginzberg. Lotus Notes® and Collaboration: Plus ça change... *Journal of Management Information Systems*, 13(3):65–81, December 1996. ISSN 0742-1222. DOI: 10.1080/07421222.1996.11518134. URL <https://doi.org/10.1080/07421222.1996.11518134>. Publisher: Routledge_eprint: <https://doi.org/10.1080/07421222.1996.11518134>.
- Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael Bernstein, and Ranjay Krishna. Explanations Can Reduce Overreliance on AI Systems During Decision-Making, December 2022. URL <http://arxiv.org/abs/2212.06823>. arXiv:2212.06823 [cs].
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.
- Jennifer Wortman Vaughan and H. Wallach. A Human-Centered Agenda for Intelligible Machine Learning. In *undefined*. 2020. URL </paper/A-Human-Centered-Agenda-for-Intelligible-Machine-Vaughan-Wallach/bc89a6fbf43cf911f71e5428d0b4a70fa5a40be9>.
- Briana Vecchione, Karen Levy, and Solon Barocas. Algorithmic Auditing and Social Justice: Lessons from the History of Audit Studies. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9, – NY USA, October 2021. ACM. ISBN 978-1-4503-8553-4. DOI: 10.1145/3465416.3483294. URL <https://dl.acm.org/doi/10.1145/3465416.3483294>.
- Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. In *CSCW 2021 - The 24th ACM Conference on Computer-Supported Cooperative Work and Social Computing*, volume 5, October 2021. DOI: 10.1145/3476068. URL <https://hal.sorbonne-universite.fr/hal-03280969>. Issue: CSCW2.
- Iris Vessey and Dennis Galletta. Cognitive Fit: An Empirical Study of Information Acquisition. *Information Systems Research*, 2(1):63–84, March 1991. ISSN 1047-7047. DOI: 10.1287/isre.2.1.63. URL <https://pubsonline.informs.org/doi/abs/10.1287/isre.2.1.63>. Publisher: INFORMS.
- Marco Virgolin, Andrea De Lorenzo, Francesca Randone, Eric Medvet, and Mattias Wahde. Model learning with personalized interpretability estimation (ML-PIE). In *Proceedings of the Genetic and Evolutionary Computation Conference Companion, GECCO '21*, pages 1355–1364, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8351-6. DOI: 10.1145/3449726.3463166. URL <https://doi.org/10.1145/3449726.3463166>.
- Sophie von Stumm, Benedikt Hell, and Tomas Chamorro-Premuzic. The hungry mind: Intellectual curiosity is the third pillar of academic performance. *Perspectives on Psychological Science*, 6(6):574–588, 2011. ISSN 1745-6924. DOI: 10.1177/1745691611421204. Place: US Publisher: Sage Publications.
- Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2):76–99, May 2017. ISSN 2044-3994. DOI: 10.1093/idpl/ix005. URL <https://doi.org/10.1093/idpl/ix005>.

- Ari Ezra Waldman. Cognitive biases, dark patterns, and the ‘privacy paradox’. *Current Opinion in Psychology*, 31:105–109, February 2020. ISSN 2352-250X. DOI: 10.1016/j.copsyc.2019.08.025. URL <https://www.sciencedirect.com/science/article/pii/S2352250X19301484>.
- Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI ’19, pages 1–15, New York, NY, USA, May 2019a. Association for Computing Machinery. ISBN 978-1-4503-5970-2. DOI: 10.1145/3290605.3300831. URL <https://doi.org/10.1145/3290605.3300831>.
- Junpeng Wang, Liang Gou, Han-Wei Shen, and Hao Yang. DQNViz: A Visual Analytics Approach to Understand Deep Q-Networks. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):288–298, January 2019b. ISSN 1077-2626, 1941-0506, 2160-9306. DOI: 10.1109/TVCG.2018.2864504. URL <https://ieeexplore.ieee.org/document/8454905/>.
- Mark Weber, Jie Chen, Toyotaro Suzumura, Aldo Pareja, Tengfei Ma, Hiroki Kanezashi, Tim Kaler, Charles E. Leiserson, and Tao B. Schardl. Scalable Graph Learning for Anti-Money Laundering: A First Look, November 2018. URL <http://arxiv.org/abs/1812.00076>. arXiv:1812.00076 [cs].
- Patrick Weber, K. Valerie Carl, and Oliver Hinz. Applications of Explainable Artificial Intelligence in Finance—a systematic review of Finance, Information Systems, and Computer Science literature. *Management Review Quarterly*, 73(1):41, February 2023. ISSN 2198-1639. DOI: 10.1007/s11301-023-00320-0. URL <https://doi.org/10.1007/s11301-023-00320-0>.
- Lisa Webley. Qualitative Approaches to Empirical Legal Research. In Peter Cane and Herbert M. Kritzer, editors, *The Oxford Handbook of Empirical Legal Research*, page 0. Oxford University Press, Oxford, November 2010. ISBN 978-0-19-954247-5. DOI: 10.1093/oxfordhb/9780199542475.013.0039. URL <https://doi.org/10.1093/oxfordhb/9780199542475.013.0039>.
- Jane Webster and Richard T. Watson. Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, 26(2):xiii–xxiii, 2002. ISSN 0276-7783. URL <https://www.jstor.org/stable/4132319>. Publisher: Management Information Systems Research Center, University of Minnesota.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, January 2023. URL <http://arxiv.org/abs/2201.11903>. arXiv:2201.11903 [cs].
- K. Weitz, D. Schiller, R. Schlagowski, T. Huber, and E. André. “Let me explain!”: exploring the potential of virtual agents in explainable AI interaction design. *Journal on Multimodal User Interfaces*, 15(2):87–98, 2021. DOI: 10.1007/s12193-020-00332-0.
- Daniel S. Weld and Gagan Bansal. The Challenge of Crafting Intelligible Intelligence. *arXiv:1803.04263 [cs]*, October 2018. URL <http://arxiv.org/abs/1803.04263>. arXiv: 1803.04263.
- Michael R. Wick and William B. Thompson. Reconstructive expert system explanation. *Artificial Intelligence*, 54(1):33–70, March 1992. ISSN 0004-3702. DOI: 10.1016/0004-3702(92)90087-E. URL <https://www.sciencedirect.com/science/article/pii/000437029290087E>.
- Christopher D. Wickens, Stephen Rice, David Keller, Shaun Hutchins, Jamie Hughes, and Krisstal Clayton. False Alerts in Air Traffic Control Conflict Alerting System: Is There a “Cry Wolf” Effect? *Human Factors*, 51(4):446–462, August 2009. ISSN 0018-7208. DOI: 10.1177/0018720809344720. URL <https://doi.org/10.1177/0018720809344720>. Publisher: SAGE Publications Inc.
- Darcia Wilkinson, Öznur Alkan, Q. Vera Liao, Massimiliano Mattetti, Inge Vejsbjerg, Bart P. Knijnenburg, and Elizabeth Daly. Why or Why Not? The Effect of Justification Styles on Chatbot Recommendations. *ACM Transactions on Information Systems*, 39(4):1–21, October 2021. URL <https://dl.acm.org/doi/10.1145/3441715>.
- Leland Wilkinson. The Grammar of Graphics: Introduction. In *The Grammar of Graphics*, Statistics and Computing, pages 1–19. Springer, New York, NY, 2005. ISBN 978-0-387-28695-2. URL https://doi.org/10.1007/0-387-28695-0_1.

- Joseph J. Williams and Tania Lombrozo. The Role of Explanation in Discovery and Generalization: Evidence From Category Learning. *Cognitive Science*, 34(5):776–806, 2010. ISSN 1551-6709. DOI: 10.1111/j.1551-6709.2010.01113.x. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1551-6709.2010.01113.x>. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1551-6709.2010.01113.x>.
- Jeannette M. Wing. Trustworthy AI. *Communications of the ACM*, 64(10):64–71, October 2021. ISSN 0001-0782, 1557-7317. DOI: 10.1145/3448248. URL <https://dl.acm.org/doi/10.1145/3448248>.
- Terry Winograd and Fernando Flores. *Understanding Computers and Cognition: A New Foundation for Design*. Addison-Wesley, 1987.
- Christine T. Wolf. Explainability scenarios: towards scenario-based XAI design. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 252–257, Marina del Ray California, March 2019. ACM. ISBN 978-1-4503-6272-6. DOI: 10.1145/3301275.3302317. URL <https://dl.acm.org/doi/10.1145/3301275.3302317>.
- Robert Wolfe. Does sunshine make a difference? *Handbook of Global Economic Governance*, 2013. Publisher: Routledge.
- Claire Woodcock, Brent Mittelstadt, Dan Busbridge, and Grant Blank. The Impact of Explanations on Layperson Trust in Artificial Intelligence-Driven Symptom Checker Apps: Experimental Study. *Journal of Medical Internet Research*, 23(11):e29386, November 2021. ISSN 1438-8871. DOI: 10.2196/29386.
- Brenda Wright. Chapter 17 - Audits and Inspections. In Delva Shamley and Brenda Wright, editors, *A Comprehensive and Practical Guide to Clinical Trials*, pages 181–183. Academic Press, January 2017. ISBN 978-0-12-804729-3. DOI: 10.1016/B978-0-12-804729-3.00017-1. URL <https://www.sciencedirect.com/science/article/pii/B9780128047293000171>.
- Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. *arXiv preprint arXiv:2101.00288*, 2021.
- Yaniv Yacoby, Ben Green, Christopher L. Griffin Jr., and Finale Doshi Velez. "If it didn't happen, why would I change my decision?": How Judges Respond to Counterfactual Explanations for the Public Safety Assessment, August 2022. URL <http://arxiv.org/abs/2205.05424>. arXiv:2205.05424 [cs].
- Jing Nathan Yan, Ziwei Gu, Hubert Lin, and Jeffrey M. Rzeszotarski. Silva: Interactively Assessing Machine Learning Fairness Using Causality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–13, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 978-1-4503-6708-0. DOI: 10.1145/3313831.3376447. URL <https://doi.org/10.1145/3313831.3376447>.
- Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, and Li Yuan. LLM Lies: Hallucinations are not Bugs, but Features as Adversarial Examples, October 2023a. URL <http://arxiv.org/abs/2310.01469>. arXiv:2310.01469 [cs].
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models, March 2023b. URL <http://arxiv.org/abs/2210.03629>. arXiv:2210.03629 [cs].
- Ji Soo Yi, Youn ah Kang, John Stasko, and J.A. Jacko. Toward a Deeper Understanding of the Role of Interaction in Information Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13(6): 1224–1231, November 2007. ISSN 1941-0506. DOI: 10.1109/TVCG.2007.70515. Conference Name: IEEE Transactions on Visualization and Computer Graphics.
- Yvette D. Clarke. Algorithmic Accountability Act of 2023, September 2023. URL <https://www.govinfo.gov/app/details/BILLS-118hr5628ih>. Call Number: Y 1.6.; Y 1.4/6; Committee: Committee on Energy and Commerce Publisher: U.S. Government Publishing Office Source: DGPO.

- Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks, November 2013. URL <http://arxiv.org/abs/1311.2901>. arXiv:1311.2901 [cs].
- John Zerilli, Alistair Knott, James Maclaurin, and Colin Gavaghan. Algorithmic Decision-Making and the Control Problem. *Minds and Machines*, 29(4):555–578, December 2019. ISSN 1572-8641. DOI: 10.1007/s11023-019-09513-7. URL <https://doi.org/10.1007/s11023-019-09513-7>.
- Baobao Zhang. Public Opinion toward Artificial Intelligence. In Justin B. Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew M. Young, and Baobao Zhang, editors, *The Oxford Handbook of AI Governance*, page o. Oxford University Press, October 2021. ISBN 978-0-19-757932-9. DOI: 10.1093/oxfordhb/9780197579329.013.36. URL <https://doi.org/10.1093/oxfordhb/9780197579329.013.36>.
- Xiaoge Zhang, Felix T. S. Chan, and Sankaran Mahadevan. Explainable machine learning in image classification models: An uncertainty quantification perspective. *Knowledge-Based Systems*, 243:108418, May 2022. ISSN 0950-7051. DOI: 10.1016/j.knosys.2022.108418. URL <https://www.sciencedirect.com/science/article/pii/S095070512200168X>.
- Zelun Tony Zhang, Yuanting Liu, and Heinrich Hussmann. Forward Reasoning Decision Support: Toward a More Complete View of the Human-AI Interaction Design Space. In *CHIItaly 2021: 14th Biannual Conference of the Italian SIGCHI Chapter*, pages 1–5, Bolzano Italy, July 2021. ACM. ISBN 978-1-4503-8977-8. DOI: 10.1145/3464385.3464696. URL <https://dl.acm.org/doi/10.1145/3464385.3464696>.
- Xun Zhao, Yanhong Wu, Dik Lun Lee, and Weiwei Cui. iForest: Interpreting Random Forests via Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):407–416, January 2019. ISSN 1941-0506. DOI: 10.1109/TVCG.2018.2864475. URL <https://ieeexplore.ieee.org/document/8454906>. Conference Name: IEEE Transactions on Visualization and Computer Graphics.
- Sarah Zheng and Jane Zhang. China Tries to Balance State Control and State Support of AI. *TIME*, August 2023. URL <https://time.com/6304831/china-ai-regulations/>.
- Joyce Zhou and Thorsten Joachims. How to Explain and Justify Almost Any Decision: Potential Pitfalls for Accountability in AI Decision-Making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23*, pages 12–21, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. DOI: 10.1145/3593013.3593972. URL <https://dl.acm.org/doi/10.1145/3593013.3593972>.
- J. Zhu, A. Liapis, S. Risi, R. Bidarra, and G. M. Youngblood. Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, pages 1–8, August 2018. DOI: 10.1109/CIG.2018.8490433. ISSN: 2325-4289.
- Alexandra Zytek, Dongyu Liu, Rhema Vaithianathan, and Kalyan Veeramachaneni. Sibyl: Understanding and Addressing the Usability Challenges of Machine Learning In High-Stakes Decision Making. *arXiv:2103.02071 [cs]*, September 2021. URL <http://arxiv.org/abs/2103.02071>. arXiv: 2103.02071.

Titre : Confiance déplacée dans l'IA : le paradoxe de l'explication et l'approche centrée sur l'homme. Une caractérisation des défis cognitifs pour faire confiance de manière appropriée aux décisions algorithmiques et applications dans le secteur financier.

Mots clés : explicabilité, apprentissage automatique, approche centrée sur l'humain, lutte anti-blanchiment, robo-advisor, supervision financière

Résumé : L'IA devenant de plus en plus présente dans nos vies, nous sommes soucieux de comprendre le fonctionnement de ces structures opaques. Pour répondre à cette demande, le domaine de la recherche en explicabilité (XAI) s'est considérablement développé au cours des dernières années. Cependant, peu de travaux ont étudié le besoin en explicabilité des régulateurs ou des consommateurs à la lumière d'exigences légales en matière d'explications. Cette thèse s'attache à comprendre le rôle des explications pour permettre la conformité réglementaire des systèmes améliorés par l'IA dans des applications financières. La première partie passe en revue le défi de prendre en compte les biais cognitifs de l'homme dans les explications des systèmes d'IA. L'analyse fournit plusieurs pistes pour mieux aligner les solutions d'explicabilité sur les processus cognitifs des individus, notamment en concevant des explications plus interactives. Elle présente ensuite

une taxonomie des différentes façons d'interagir avec les solutions d'explicabilité. La deuxième partie se concentre sur des contextes financiers précis. Une étude porte sur les systèmes de recommandation et de souscription en ligne de contrats d'assurance-vie. L'étude souligne que les explications présentées dans ce contexte n'améliorent pas de manière significative la compréhension de la recommandation par les utilisateurs non experts. Elles ne suscitent pas davantage la confiance des utilisateurs que si aucune explication n'était fournie. Une autre étude analyse les besoins des régulateurs en matière d'explication dans le cadre de la lutte contre le blanchiment d'argent et le financement du terrorisme. Elle constate que les autorités de contrôle ont besoin d'explications pour établir le caractère répréhensible des cas de défaillance échantillonnés, ou pour vérifier et contester la bonne compréhension de l'IA par les banques.

Title : Misplaced trust in AI: the explanation paradox and the human-centric path. A characterisation of the cognitive challenges to appropriately trust algorithmic decisions and applications in the financial sector.

Keywords : explainability, machine learning, human-centered approach, anti-money laundering, robo-advisor, financial supervision

Abstract : As AI is becoming more widespread in our everyday lives, concerns have been raised about comprehending how these opaque structures operate. In response, the research field of explainability (XAI) has developed considerably in recent years. However, little work has studied regulators' need for explainability or considered effects of explanations on users in light of legal requirements for explanations. This thesis focuses on understanding the role of AI explanations to enable regulatory compliance of AI-enhanced systems in financial applications. The first part reviews the challenge of taking into account human cognitive biases in the explanations of AI systems. The analysis provides several directions to better align explainability solutions with people's cognitive processes, including designing

more interactive explanations. It then presents a taxonomy of the different ways to interact with explainability solutions. The second part focuses on specific financial contexts. One study takes place in the domain of online recommender systems for life-insurance contracts. The study highlights that feature-based explanations do not significantly improve non expert users' understanding of the recommendation, nor lead to more appropriate reliance compared to having no explanation at all. Another study analyzes the needs of regulators for explainability in anti-money laundering and financing of terrorism. It finds that supervisors need explanations to establish the reprehensibility of sampled failure cases, or to verify and challenge banks' correct understanding of the AI.