The Explanation Scope Does not Fit all: Local, Model-Centric and the Role of Cognitive Traits

Federico Maria Cau
University of Cagliari
Cagliari, Italy
federicom.cau@unica.it

Lucio Davide Spano
University of Cagliari
Cagliari, Italy
davide.spano@unica.it

Abstract

Explainable AI (XAI) aims to support human decision-making by improving understanding and fostering calibrated trust. Yet, it remains unclear whether specific explanation types consistently help users make better decisions, and how user traits such as Need for Cognition (NFC) influence their effects. We present a confirmatory analysis of two controlled user studies in different domains (loan approval and job candidate screening), comparing local, feature-based, and global, model-centric explanations. We analyze decision accuracy and over-reliance as a function of AI confidence and correctness, while accounting for individual differences in NFC.

Across both tasks, AI confidence emerged as the strongest predictor of human accuracy: users were significantly more likely to follow correct AI recommendations when confidence was high. Local explanations further boosted accuracy on correct predictions. When the AI was wrong and low-confident, explanation effects varied by user trait: local explanations reduced over-reliance among low-NFC participants but had the opposite effect for high-NFC individuals. These results highlight that explanation effectiveness depends on model correctness, user traits, and context. We conclude with design implications for confidence-aware, trait-sensitive XAI systems that adapt explanation delivery to user profiles and prediction uncertainty.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; **Interaction design**; • **Computing methodologies** → **Artificial intelligence**.

Keywords

AI-Assisted Decision-making, Explainable AI, Model-centric Explanations, Need for Cognition

ACM Reference Format:
Federico Maria Cau and Lucio Davide Spano. 2025. The Explanation Scope Does not Fit all: Local, Model-Centric and the Role of Cognitive Traits. In *CHItaly 2025: 16th Biannual Conference of the Italian SIGCHI Chapter (CHItaly 2025), October 06–10, 2025, Salerno, Italy*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3750069.3750343

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
CHItaly 2025, Salerno, Italy
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2102-1/25/10
https://doi.org/10.1145/3750069.3750343

1 Introduction

Artificial Intelligence (AI) systems are increasingly deployed to assist human decision-making in high-stakes domains such as finance, recruitment, and healthcare. A central promise of explainable AI (XAI) is to provide insights into algorithmic decisions, helping users make more accurate and fair judgments while avoiding over-reliance on automated suggestions [1]. Despite this promise, the real-world benefits of different explanation types remain a matter of contention. In particular, it is unclear whether users benefit more from specific *explanation scopes* such as local or instance-level explanations, versus model-centric summaries or other interface factors, including the AI model's expressed *confidence* in its predictions.

In this paper, we investigate how explanation scope and model confidence affect user decisions in two AI-assisted scenarios involving consequential judgments: loan approvals and job candidate screening. Both scenarios are prototypical high-stakes tasks involving structured tabular data, where AI and humans team together. Our analysis focuses on interaction conditions in which the AI provides its prediction and the estimated confidence, together with either (i) local feature-based explanations or (ii) model-centric summaries (e.g., rules or global indicators). Our research questions are the following:

RQ1 How do explanation scope and model confidence influence decision accuracy and overreliance in AI-assisted high-stakes tasks?

RQ2 Do individual differences in Need for Cognition reduce over-reliance on incorrect AI predictions through explanations?

Prior work has suggested that local feature-based explanations (e.g., SHAP [26]) can enhance user understanding of specific predictions, especially when paired with well-calibrated confidence estimates [15, 21]. However, others have found that such explanations may increase cognitive load or reinforce overreliance, especially when the model is wrong [16]. Global or model-centric summaries may be less cognitively demanding but offer weaker support for task-level reasoning [18, 19, 24].

In addition to interface-level factors, we also consider whether individual cognitive traits affect user behavior. Specifically, we examine *Need for Cognition* [14] (NFC), a stable psychological trait that reflects a person's tendency to engage in and enjoy effortful cognitive activities. Prior work has suggested that users with higher NFC may be more inclined to scrutinize AI recommendations and resist blind reliance, particularly in ambiguous or error-prone scenarios [2, 10, 23, 28].

To explore this space, we evaluate two hypotheses: local explanation coupled with high model confidence leads to higher human decision accuracy than a model-centric explanation (H1), and on trials where the model is wrong and expresses low confidence, the

need for cognition (NFC) traits moderates the effectiveness of the explanation scope. We test these hypotheses with consistent interface designs and calibrated model outputs across two tasks. Our results show that explanation effects are most beneficial when paired with high-confidence, correct model predictions, and that the reliance of participants with low NFC benefit from local explanations.

Our contribution is twofold. We (1) present an empirical comparison of explanation scope and model confidence effects across tasks, and (2) discuss implications for building more adaptive, user-centered decision support systems.

2 Related Work

Explainable AI (XAI) aims to improve human understanding of AI-generated decisions, particularly in high-stakes domains where trust, accuracy, and appropriate reliance are critical. However, the literature reveals persistent disagreements over which explanation types are helpful, how model confidence should be communicated, and how individual cognitive traits shape the impact of AI explanations. This section synthesizes previous work on these aspects, drawing from studies that examined the effects of explanation styles, model confidence, and personality traits on decision-making within human-AI teams.

2.1 Explanation Style in XAI

The effectiveness of different explanation styles in improving user outcomes is a core question in XAI. Most empirical work has focused on local, feature-based techniques such as SHAP or LIME, which highlight the contribution of input features to a specific prediction [2, 3]. Some studies find that local explanations improve decision accuracy when the AI is correct [2, 4], but others suggest they can increase overreliance, especially when the model is wrong [3, 6].

Alternative styles, such as rule-based [11] and counterfactual explanations [27], are often proposed to enhance interpretability, but their comparative effectiveness remains underexplored. Cau and Spano [18] found that counterfactual explanations reduced cognitive load and improved accuracy when the AI was correct, despite being rated as harder to understand. Feature-based explanations, by contrast, were more familiar to users but did not significantly boost performance. These findings add to growing evidence that the benefits of specific explanation styles are highly context-dependent and may not generalize across tasks or users [2, 10].

2.2 Communicating Confidence and Calibrating Trust

Another key factor in XAI is how model confidence is communicated. Providing AI confidence scores has been shown to influence user trust, sometimes more than the model's accuracy [3, 6]. Users tend to follow high-confidence predictions even without strong accuracy guarantees. Cau and Spano [18] found that confidence alone substantially impacted user behaviour, increasing reliance while reducing cognitive load. In contrast, when confidence was low, users were more likely to engage in analytical reasoning and reject AI suggestions.

These findings align with prior work suggesting that AI confidence can act as a "trust anchor" [9], especially when users must quickly assess the reliability of the model. However, overly confident AI outputs can mislead users into accepting incorrect predictions, raising the need for well-calibrated and transparent confidence communication [8, 22]. Balancing informativeness with caution in how confidence is presented remains a key challenge for XAI design.

2.3 Overreliance and Task Context

A recurring issue in human-AI collaboration is overreliance: users accepting AI suggestions even when the model is wrong. Prior studies have shown that explanation style and interface design can influence overreliance, but task framing and perceived stakes also play a large role [7, 25].

In contrast, using an on-demand explanation paradigm, no significant reduction in overreliance across explanation types was found [19]. The results suggest that giving users the option to request explanations might reduce automatic compliance, but this effect depends on the user's cognitive engagement and the clarity of the information presented.

2.4 Need for Cognition and User Traits

Need for Cognition (NFC) is a stable personality trait that reflects an individual's motivation to engage in and enjoy analytical thinking [14]. High-NFC individuals are typically more reflective and less prone to heuristic shortcuts, leading to the hypothesis that they might benefit more from detailed explanations in AI-assisted decision-making. However, empirical findings on this relationship are mixed.

Prior work found that high-NFC individuals might benefit from conditions that trigger deeper cognitive processing, such as cognitive forcing interventions [10]. For instance, showing explanations only after an initial decision or requiring users to request them on demand may better leverage individual differences in NFC. Nevertheless, recent work [12, 13, 19] highlights that cognitive intervention might not be enough to reduce overreliance on AI for individuals with high-NFC.

For example, Cau and Spano [18] found no main effect of NFC on decision accuracy or cognitive load but did observe some differences in how low and high NFC individuals prioritized interface elements, while in [19] NFC showed positive correlations with other curiosity-related measures (e.g., Epistemic Curiosity, CEI-II) and was associated with greater self-reported confidence. However, NFC did not reduce overreliance, and high-NFC users did not perform better when exposed to complex or hybrid on-demand explanations.

3 Methods

To investigate the effects of explanation scope and model confidence on human-AI decision-making, we analysed the data available from two controlled studies focused on AI-assisted decisions in high-stakes settings: loan approval and job candidates selection [18, 19]. Both studies share a common structure: users receive instances to decide on (i.e., a loan or job applications), and a prediction from a calibrated AI system. Experimental conditions vary the explanation scope (local or model-centric) and AI confidence levels (high or

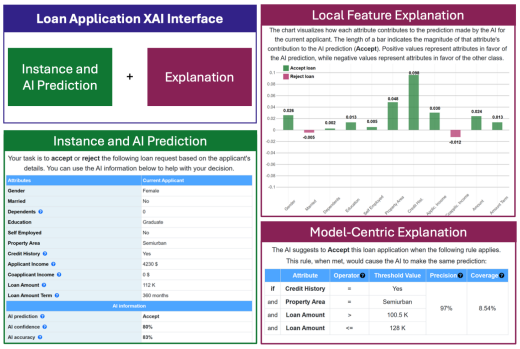

Figure 1: XAI interface for loan application decision. The interface consists of two parts: the instance and AI prediction group, where the interface presents information about the current application and the AI suggested decision, and the explanation group, which could include either local feature (top part) or model-centric explanations (bottom part).

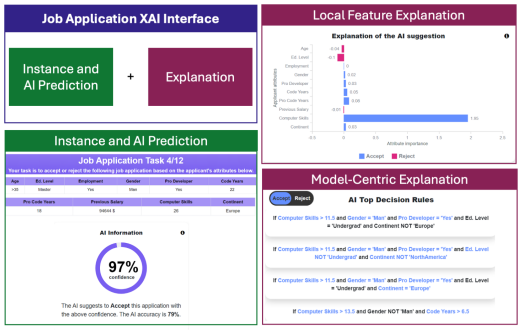

Figure 2: XAI interface for job application decision. The interface organization is the same as in Figure 1.

low). The participant's task is to decide whether to accept or reject a given job or loan application.

By analyzing two distinct tasks, we aim to strengthen the external validity of these findings. Each task reflects a realistic, high-stakes decision scenario with structured input data and varying class imbalance. While the tasks differ in context, their structure is consistent: both present a series of binary classification decisions, with users supported by AI predictions and explanations, under experimental control of explanation scope and model confidence.

3.1 Loan Approval Task

The loan approval task was based on the publicly available Loan Prediction dataset[1], which includes 614 real-world loan applications. Each instance describes an applicant using twelve attributes (e.g., credit history, income, loan amount). We trained a Random Forest Classifier (RFC) with 100 estimators using an 80:20 stratified train-test split, reaching 83% test accuracy. Model confidence scores were derived from Shannon entropy-based uncertainty estimates, scaled to a 0–100 range. Instances were categorized as low-confidence (<44.3) or high-confidence (>61.6) based on the lower and upper quartiles of the confidence distribution.

For the study, participants were shown eight AI-assisted instances (plus a separate practice set), balanced by AI correctness (correct vs. wrong), confidence (low vs. high), and predicted class. The user interface displayed either a **feature-based explanation** (via SHAP attributions) or a **rule-based explanation** (via Anchor rules), alongside the AI prediction and confidence. Each participant was assigned to one explanation condition in a between-subjects design. The study included attention checks and a monetary incentive scheme (£0.12 per correct answer) to encourage engagement and simulate real-world stakes. Figure 1 shows an example decision instance.

[1]https://www.kaggle.com/datasets/altruistdelhite04/loan-prediction-problem-dataset

3.2 Job Candidate Screening Task

The other task involved evaluating job applicants based on structured profiles derived from a preprocessed version of a large job application dataset. Each application consisted of ten features, including age, education, and years of programming experience. We trained and calibrated an Extreme Gradient Boosting (XGB) model using a 60:20:20 train-calibration-test split, achieving test accuracy around 78.5%. Confidence levels were computed from calibrated output probabilities and binned into three categories; for our analysis, we only retained **high** (>0.8) and **low** (<0.55) confidence instances.

The study used a between-subjects design with four explanation conditions. For our purposes, we isolate two: **local model-centric** explanations (akin to SHAP-style feature contributions) and **global model-centric** explanations (summarized through global rule sets or model behaviour descriptions). Each participant completed ten main decision trials, preceded by a practice session. Figure 2 shows an example decision instance.

3.3 Common Procedure and Measures

Both studies followed a similar structure. Participants were recruited online through the Prolific platform and completed attention checks to ensure data quality. Tasks were implemented in LimeSurvey, with consistent timing, compensation rates (£2.7 base + bonuses), and incentive framing. Participants were randomly assigned to one explanation condition, with instances controlled for balance across correctness and confidence. We measured NFC using the NCS-6 six-item five-point scale (1 = extremely uncharacteristic of me; 5 = extremely characteristic of me) as defined in [20], assigning participants to low or high levels by comparing their score to the distribution median.

For each trial, we recorded participants' decisions, AI correctness, confidence level, and explanation condition. Across tasks, dependent variables include: (1) **decision accuracy**, whether the user aligned with the ground truth; and (2) **overreliance**, measured as agreement with incorrect AI predictions. In both studies, confidence and correctness were manipulated within-subjects, while explanation scope was a between-subjects factor.

3.4 Sample Size and Hypothesis Testing Strategy

We designed this analysis to test two pre-specified hypotheses regarding the influence of explanation scope, model confidence, and individual cognitive traits on user decision quality and overreliance in AI-assisted tasks. These hypotheses were grounded in prior work on explainable AI, trust calibration, and cognitive motivation [11, 14, 15]:

- **H1.** *When the AI prediction is correct, both explanation scope and model confidence will have independent effects on user decision accuracy.*
- **H2.** *When the AI prediction is wrong and expresses low confidence, Need for Cognition (NFC) moderates the effect of explanation scope on overreliance.*

To test these hypotheses, we reanalyzed data from two previously conducted user studies involving AI-assisted decision-making. We selected only participants from conditions that presented either local or model-centric explanations and excluded hybrid, counterfactual, and no-AI conditions to minimize interface variation. This resulted in 96 participants from the job application screening task [19] and 146 participants from the loan approval task [18], for a total of 242 participants.

We tested each hypothesis using separate mixed-effects logistic regression models:

- For **H1**, we restricted the data to trials where the AI prediction was *correct* and modelled decision accuracy (1 = correct, 0 = incorrect) as a function of explanation scope (local vs. model-centric), model confidence (high vs. low), and their interaction.
- For **H2**, we restricted the data to trials where the AI prediction was *incorrect* and modelled overreliance (1 = copied incorrect AI, 0 = disagreed) as a function of NFC, explanation scope, and model confidence.

Both models included random intercepts for participants to account for repeated measures. Predictors were effect-coded, and statistical significance was assessed using likelihood ratio tests at $\alpha = .05$. We report coefficient estimates, odds ratios, and 95% confidence intervals to support interpretation.

4 Results

4.1 H1: Decision accuracy with correct, high confident AI predictions and local explanations

To test H1, we considered only trials in which the classifier's prediction was *correct* and the reported confidence was *low* or *high* The resulting analysis set contains $N = 1\,066$ decisions (472 from the loan-approval task, 594 from the job-screening task).

We fitted a logistic model with participant accuracy (1 = correct) as the dependent variable:

$$accuracy \sim scope \times confidence,$$

where *confidence* = 1 for "high" and 0 for "low", and *scope* = 1 for the feature-based explanation (SHAP/sXAI) and 0 for the model-centric alternative (Anchor/MCE[2]). Table 1 shows the results.

[2]Model-Centric Explanation (MCE)

Table 1: Logistic regression on AI-correct trials (*med* confidence excluded).

Predictor	β	SE	OR	p
Intercept (low, model-centric)	−0.03	0.11	0.97	.77
High confidence	**+1.04**	0.18	**2.83**	**<.001**
Local explanation	**+0.38**	0.17	**1.46**	**.030**
Local × High conf.	+0.06	0.18	1.06	.73

Table 2: Logit model on AI-wrong, low-confidence trials (binary NFC tails).

Predictor	β	SE	OR	p
Intercept (hiring, model, low NFC)	+1.44	0.48	4.22	.003
Local explanation	−0.94	0.58	0.39	.10
High NFC	−0.74	0.59	0.48	.21
Local × High NFC	**+1.64**	0.80	**5.16**	**.041**

Moving from a low-confidence to a high-confidence prediction nearly *triples* the odds that the human decision is correct (OR ≈ 2.8). Independent of the confidence effect, a local, feature-based explanation increases accuracy by about 6% (odds ratio = 1.46) compared with a model-centric summary. The interaction term is small and non-significant, indicating that the local-explanation advantage holds across the binary confidence split. Taken together, these results **support H1**: when the AI is correct, combining high confidence with a local explanation yields the most reliable human–AI performance.

4.2 H2: Overreliance when the AI is wrong & low-confident

We retained only trials in which the classifier was *wrong* and declared *low* confidence, yielding $N = 133$ decisions. Overreliance was coded 1 whenever the participant copied the AI's incorrect answer. The logistic model was

$$overreliance \sim scope \times nfc$$

where $local = 1$ for feature-based explanations and nfc_z is the standardised Need-for-Cognition score. Table 2 reports the coefficients.

The significant interaction ($\beta = 1.64,\ p = .041$) confirms that NFC *moderates* how explanation scope affects overreliance, **supporting H2**. The moderation is a *crossover*: for low-NFC participants, a local explanation reduces over-reliance (from roughly 28 % with a model-centric summary to 11 %), but for high-NFC participants the same local explanation *increases* copying (from 17 % to about 24 %). In other words, feature-based explanations help less analytical users resist an uncertain, wrong AI while inadvertently persuading analytical users to follow it.

4.3 Summary

H1 was supported: when the AI prediction was correct, decision accuracy increased with local explanations and high confidence.

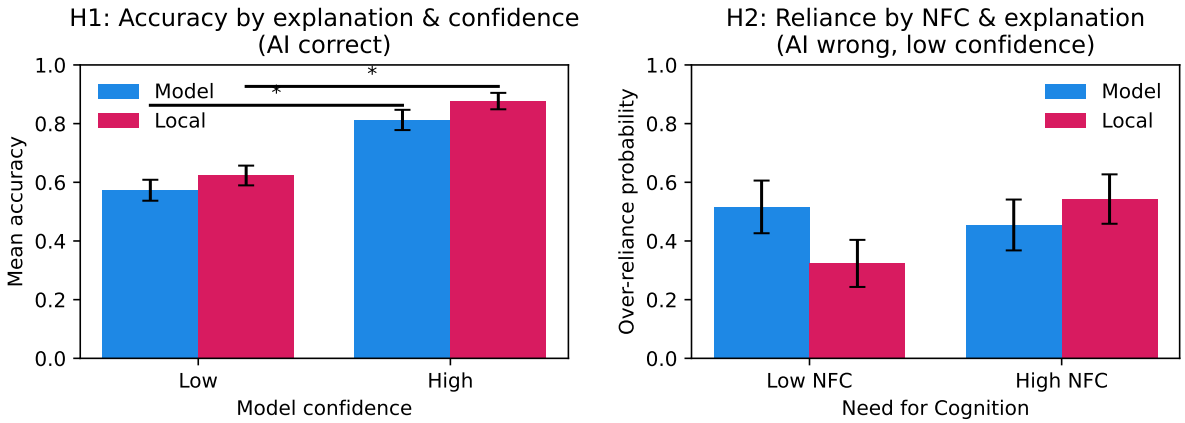

Figure 3: Overall effects of confidence, explanation scope, and Need for Cognition (NFC). On the left, we show the decision accuracy on trials where the AI prediction was *correct*. Bars compare model-centric (blue) and local, feature-based explanations (magenta) at *low* and *high* model-confidence levels. On the right, we show the overreliance rate—probability of copying the AI's recommendation—on trials where the AI was *wrong* and *low-confident*, split by *low* vs *high* NFC.

H2 was supported: NFC moderates the explanation scope with a crossover interaction: low NFC participants benefit from local explanations, while for high NFC participants, the same explanations increase overreliance.

5 Discussion

Local explanations and confidence boost accuracy when AI is correct. Our findings support H1: in trials where the AI was correct, users were more likely to make correct decisions when supported by high model confidence and local, feature-based explanations. This confirms prior research showing that instance-specific explanations can align human and model reasoning, improving decision accuracy in high-stakes tasks [2, 4]. While model-centric summaries offer structured insights into global logic, they may lack the specificity users need when evaluating a single instance.

In addition, AI confidence served as a strong cue for the final outcome: participants consistently performed better when the AI expressed high certainty. This confirms past findings on the effect of a high AI confidence in guiding user trust more than explanation content [3, 6, 17]. The additive effects observed here suggest that explanation scope and confidence play a meaningful role in helping users align with correct AI predictions.

Overreliance patterns depend on explanation scope and NFC. Our second hypothesis, H2, was also supported. We observed a significant crossover interaction: participants with low Need for Cognition (NFC) were less likely to overrely on incorrect, low-confidence AI predictions when given local explanations. In contrast, high-NFC participants showed increased overreliance under the same condition.

One possible interpretation is that low-NFC users benefit from the concreteness of local explanations, which reduce ambiguity and discourage blind copying. This matches prior observations that feature-based explanations help users stay grounded when AI predictions are uncertain [25]. Conversely, high-NFC users may engage more deeply with the explanation content, leading to greater rationalization of the AI's prediction, even when wrong. This effect echoes concerns raised in [18, 25], where explanation richness sometimes backfired by reinforcing the perceived plausibility of incorrect predictions.

Implications for explanation design and personalization. Our results underscore the importance of tailoring XAI interfaces to system-level and user-level factors. While confidence and explanation scope independently support decision accuracy when the AI is correct, their effects on overreliance depend on cognitive traits like NFC. This challenges the idea that more explanation is always better and supports a more adaptive approach to explanation delivery [5, 25].

One promising direction is the use of confidence-aware and user-sensitive explanation policies. Prior studies have proposed adaptive explanation timing and selective disclosure [2], particularly when the model is uncertain or the user is cognitively disengaged. Our results suggest these strategies could be improved by factoring in user traits like NFC. For instance, local explanations might be emphasized for low-NFC users but withheld or reframed for high-NFC users when model confidence is low. Further research is required to identify other traits that could foster building appropriate reliance on the AI's suggestion.

Toward trait-sensitive, context-aware XAI. The findings in this paper contribute to the current shift in XAI research from evaluating static explanation formats to designing interfaces that respond to users and context. Past studies have shown that explanation scope, user agency, and perceived stakes all shape behaviour in human-AI interaction [3, 17, 25]. However, besides the AI confidence, it is difficult to find consistent predictors to optimize the AI-user teaming accuracy. Our study adds to this literature by demonstrating that even validated user traits like NFC can reverse the effects of explanation scope under specific conditions.

Future XAI systems will need to go beyond focusing on computing metrics about relevant aspects of the instance and the AI model, or defining an appropriate visualization. They must detect when explanations are helpful, for whom, and under what conditions. Our findings suggest that combining calibrated confidence displays with adaptive explanation strategies could be a promising step toward more effective human–AI collaboration.

6 Conclusion

This paper examined how explanation scope, model confidence, and user traits affect decision quality and reliance in AI-assisted decision-making. Motivated by inconsistent findings in prior XAI literature, we analyzed two tasks, loan approval and job screening, focusing on the specific conditions under which explanations help or hinder human–AI collaboration.

Our results show that when the AI prediction is correct, combining high confidence with local, feature-based explanations leads to the highest user accuracy. However, when the AI is wrong and low-confident, explanations offer no consistent protection against overreliance. Moreover, individual differences matter: participants with low Need for Cognition benefited from local explanations when the AI was wrong and uncertain, whereas for high-NFC users, the same explanations increased the likelihood of over-reliance.

Future work should expand this investigation along several dimensions. First, replication across additional domains and explanation types—such as counterfactuals, contrastive rationales, or hybrid approaches—would clarify the generality of these effects. Second, real-time or longitudinal studies could assess how trust and reliance evolve over repeated interactions. Finally, adaptive systems that personalize explanation complexity or trigger user reflection in uncertain cases hold promise for building more robust and equitable human–AI partnerships.

Acknowledgments

This research is partially funded by the Italian Ministry of University and Research (MUR) and by the European Union - NextGenerationEU, Mission 4, Component 2, Investment 1.1, under grant PRIN 2022 PNRR "DAMOCLES: Detection And Mitigation Of Cyber attacks that exploit human vuLnerabilitiES" (Grant P2022FXP5B) — CUP: F53D23009220001.

References

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. doi:10.1109/ACCESS.2018.2870052

[2] Vedant Bahel, Harshinee Sriram, and Cristina Conati. 2024. Initial results on personalizing explanations of AI hints in an ITS. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization* (Cagliari, Italy) *(UMAP '24)*. Association for Computing Machinery, New York, NY, USA, 244–248. doi:10.1145/3627043.3659566

[3] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. 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* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. doi:10.1145/3411764.3445717

[4] Astrid Bertrand, Rafik Belloum, James R. Eagan, and Winston Maxwell. 2022. How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) *(AIES '22)*. Association for Computing Machinery, New York, NY, USA, 78–91. doi:10.1145/3514094.3534164

[5] Aditya Bhattacharya, Jeroen Ooge, Gregor Stiglic, and Katrien Verbert. 2023. Directive Explanations for Monitoring the Risk of Diabetes Onset: Introducing Directive Data-Centric Explanations and Combinations to Support What-If Explorations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 204–219. doi:10.1145/3581641.3584075

[6] Aditya Bhattacharya, Simone Stumpf, Lucija Gosak, Gregor Stiglic, and Katrien Verbert. 2024. EXMOS: Explanatory Model Steering through Multifaceted Explanations and Data Configurations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 314, 27 pages. doi:10.1145/3613904.3642106

[7] Aditya Bhattacharya, Simone Stumpf, and Katrien Verbert. 2024. An Explanatory Model Steering System for Collaboration between Domain Experts and AI. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization* (Cagliari, Italy) *(UMAP Adjunct '24)*. Association for Computing Machinery, New York, NY, USA, 75–79. doi:10.1145/3631700.3664886

[8] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Montreal QC</city>, <country>Canada</country>, </conf-loc>) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3173574.3173951

[9] Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 2022. 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* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 807–819. doi:10.1145/3490099.3511139

[10] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (April 2021), 21 pages. doi:10.1145/3449287

[11] Zana Buçinca, Max Malaya, Kathryn Glass, and Gagan Bansal. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 454–464.

[12] Zana Buçinca, Siddharth Swaroop, Amanda E. Paluch, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2024. Contrastive Explanations That Anticipate Human Misconceptions Can Improve Human Decision-Making Skills. arXiv:2410.04253 [cs.HC] https://arxiv.org/abs/2410.04253

[13] Zana Buçinca, Siddharth Swaroop, Amanda E. Paluch, Susan A. Murphy, and Krzysztof Z. Gajos. 2024. Towards Optimizing Human-Centric Objectives in AI-Assisted Decision-Making With Offline Reinforcement Learning. arXiv:2403.05911 [cs.HC] https://arxiv.org/abs/2403.05911

[14] John T Cacioppo and Richard E Petty. 1982. The need for cognition. *Journal of Personality and Social Psychology* 42, 1 (1982), 116–131. doi:10.1037/0022-3514.42.1.116

[15] Carrie J Cai, Emily Reif, Neel Hegde, James Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg Corrado, Jeff Dean, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[16] Federico Maria Cau, Hanna Hauptmann, Lucio Davide Spano, and Nava Tintarev. 2023. Effects of AI and Logic-Style Explanations on Users' Decisions Under Different Levels of Uncertainty. *ACM Trans. Interact. Intell. Syst.* 13, 4, Article 22 (Dec. 2023), 42 pages. doi:10.1145/3588320

[17] Federico Maria Cau, Hanna Hauptmann, Lucio Davide Spano, and Nava Tintarev. 2023. Supporting High-Uncertainty Decisions through AI and Logic-Style Explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 251–263. doi:10.1145/3581641.3584080

[18] Federico Maria Cau and Lucio Davide Spano. 2025. Exploring the Impact of Explainable AI and Cognitive Capabilities on Users' Decisions. arXiv:2505.01192 [cs.AI] https://arxiv.org/abs/2505.01192

[19] Federico Maria Cau and Lucio Davide Spano. 2025. The Influence of Curiosity Traits and On-Demand Explanations in AI-Assisted Decision-Making. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 1440–1457. doi:10.1145/3708359.3712165

[20] Gabriel Lins de Holanda Coelho, Paul H. P. Hanel, and Lukas J. Wolf. 2020. The Very Efficient Assessment of Need for Cognition: Developing a Six-Item Version. *Assessment* 27, 8 (2020), 1870–1885. doi:10.1177/1073191118793208 arXiv:https://doi.org/10.1177/1073191118793208 PMID: 30095000.

[21] Sander de Jong, Ville Paananen, Benjamin Tag, and Niels van Berkel. 2025. Cognitive Forcing for Better Decision-Making: Reducing Overreliance on AI Systems Through Partial Explanations. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–30.

[22] Krzysztof Z. Gajos and Krysta Chauncey. 2017. The Influence of Personality Traits and Cognitive Load on the Use of Adaptive User Interfaces. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces* (Limassol, Cyprus) *(IUI '17)*. Association for Computing Machinery, New York, NY, USA, 301–306. doi:10.1145/3025171.3025192

[23] Krzysztof Z. Gajos and Lena Mamykina. 2022. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In *Proceedings of the 27th International Conference on Intelligent User Interfaces* (Helsinki, Finland,) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 794–806. doi:10.1145/3490099.3511138

[24] Lukas-Valentin Herm. 2023. Impact of explainable ai on cognitive load: Insights from an empirical study. *arXiv preprint arXiv:2304.08861* (2023).

[25] Daniel Herzog and Wolfgang Wörndl. 2019. A User Study on Groups Interacting with Tourist Trip Recommender Systems in Public Spaces. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization* (Larnaca, Cyprus) *(UMAP '19)*. Association for Computing Machinery, New York, NY, USA, 130–138. doi:10.1145/3320435.3320449

[26] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[27] Sandra Wachter, Brent Daniel Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Cybersecurity* (2017). https://api.semanticscholar.org/CorpusID:3995299

[28] Josephine Zerna, Anja Strobel, and Alexander Strobel. 2024. The Role of Need for Cognition in Well-Being – Review and Meta-Analyses of Associations and Potentially Underlying Mechanisms. *Collabra: Psychology* 10, 1 (02 2024), 92885. doi:10.1525/collabra.92885 arXiv:https://online.ucpress.edu/collabra/article-pdf/10/1/92885/826822/collabra_2024_10_1_92885.pdf