AI Governance and Society Journal
ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100
https://galaxiauniverse.com/index.php/AIGSJ/issue/archive

# Human-Understandable Representations for Complex AI Decisions

**Laura Klein[1], Helena Klein[2], Marta Kovacs[3]**

[1] Research Scientist, Department of Artificial Intelligence, Mediterranean Institute of Technology, Rome, Italy. Email: laura.klein9@gmail.com | ORCID: 7316-9883-9873-6273
[2] Professor, Department of Artificial Intelligence, Nordic Technical University, Stockholm, Sweden. Email: helena.klein203@gmail.com | ORCID: 1540-2594-9427-2920
[3] Assistant Professor, Department of Computer Science, Baltic AI Research University, Tallinn, Estonia. Email: marta.kovacs309@gmail.com | ORCID: 3740-6727-9094-2392

## ABSTRACT

*Complex AI decision systems -- encompassing deep neural networks, large language models, and ensemble architectures -- produce outputs of considerable practical consequence yet generate internal representations that are fundamentally inaccessible to human cognition in their raw form. Human- understandable representations (HURs) are structured reformulations of model-internal information designed to align with human cognitive schemas, domain knowledge, and explanation needs, enabling meaningful human oversight and accountability for AI-assisted decisions. Despite extensive research on explainability methods, the systematic design of HURs -- as distinct from the application of generic attribution techniques -- has received limited attention. This paper proposes a HUR design framework comprising four representation classes (contrastive, causal, conceptual, and narrative) and evaluates their cognitive effectiveness through a user study involving 186 participants across three professional domains: clinical medicine (n = 62), financial advising (n = 62), and legal practice (n = 62). Participants assessed AI decisions with and without HURs across four cognitive effectiveness dimensions: comprehension accuracy, decision confidence calibration, appropriate reliance, and perceived fairness. Results demonstrate that HUR-supported conditions achieve a 41.8% improvement in comprehension accuracy (SD = 6.2%), a 33.5% improvement in decision confidence calibration (SD = 5.4%), and a 27.6% improvement in appropriate reliance (SD = 4.9%) relative to no-HUR controls. Representation class effectiveness is domain-dependent: contrastive HURs are most effective in legal contexts, causal HURs in clinical contexts, and conceptual HURs in financial contexts. The study contributes the HUR design framework, domain-specific design guidelines, and a validated HUR evaluation instrument.*

**Keywords:** human-understandable representations; AI explainability; cognitive effectiveness; contrastive explanations; causal explanations; human oversight; responsible AI; EU AI Act

**Citation:** Klein et al. [2025]. Human-Understandable Representations for Complex AI Decisions. DOI: https://doi.org/10.5281/zenodo.19184989



**Article Information:** *Received:* 22 November 2024 *Accepted:* 24 January 2025 *Published:* 28 March 2025

**Research Article:** *Research Article*

## 1. Introduction

### 1.1 Beyond Attribution: The HUR Design Challenge

The field of explainable AI has invested heavily in developing techniques that reveal which input features most strongly influence a model's predictions -- SHAP, LIME, Integrated Gradients, and GradCAM among them (Lundberg and Lee, 2017; Ribeiro et al., 2016; Sundararajan et al., 2017; Selvaraju et al., 2017). Yet a growing body of evidence from cognitive psychology and human-computer interaction research indicates that feature attribution explanations, however technically accurate, often fail to produce the cognitive outcomes that explanation is intended to achieve: accurate comprehension of the AI's reasoning, well-calibrated decision confidence, and appropriate reliance -- neither overtrusting nor undertrusting the AI's recommendations (Bansal et al., 2021; Poursabzi-Sangdeh et al., 2021; Schemmer et al., 2023). The reason, argued by Miller (2019), is that human explanatory cognition is not fundamentally attribution-based. Drawing on four decades of research in cognitive psychology and the social sciences, Miller demonstrates that humans preferentially seek contrastive explanations (why P rather than Q), causal accounts (what caused P), and social references (what would a comparable agent have done). Feature attribution maps -- lists of numerical scores for input dimensions -- are fundamentally misaligned with these cognitive preferences, explaining why empirical studies consistently find that attribution explanations improve stated satisfaction but not objective comprehension or decision quality (Poursabzi-Sangdeh et al., 2021; Chromik and Butz, 2021). Human-understandable representations (HURs) address this misalignment by reformulating model-internal information in terms of the cognitive schemas, causal frameworks, and conceptual vocabularies of the intended explanation recipient. A contrastive HUR does not report a feature attribution score but answers the question the recipient is actually asking: why was this patient classified as high-risk rather than low-risk? A causal HUR does not display a saliency map but traces the causal chain from input variables to output through a domain-relevant causal model. The design of HURs is therefore a cognitive design challenge as much as a technical one, requiring alignment between representation structure and human cognitive process.

### 1.2 Research Objectives and Paper Structure

This paper proposes a HUR design framework comprising four representation classes and evaluates their cognitive effectiveness through a large-scale user study. The research objectives are: (1) to specify and motivate four HUR classes -- contrastive, causal, conceptual, and narrative -- drawing on cognitive psychology, social science explanation theory, and AI explainability research; (2) to evaluate the cognitive effectiveness of each HUR class across three professional domains through a controlled user study (n = 186); and (3) to derive domain-specific HUR design guidelines and a validated evaluation instrument. Section 2 reviews prior work on explanation cognition, HCI research on AI explanation, and relevant regulatory requirements. Section 3 presents the HUR design framework. Section 4 describes the user study methodology. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

## 2. Literature Review

### 2.1 Cognitive Foundations of Human Explanation

The cognitive psychology of explanation has established several robust findings about how humans process and evaluate explanations. Hilton (1990) demonstrated that explanations are inherently contrastive -- humans explain events by identifying what made them occur rather than an alternative (the foil). Lombrozo (2006) showed that explanatory depth preferences interact with the purpose of the explanation: people seek mechanistic, causal explanations for novel phenomena but prefer simpler, categorisation-based explanations for familiar events. Khemlani and Johnson-Laird (2011) demonstrated that counterfactual reasoning -- a core component of both contrastive and causal explanation -- relies on mental simulation of alternative scenarios, a cognitively demanding process that can be supported or impeded by explanation format. Miller (2019) synthesised these findings in the context of AI explanation, arguing that effective AI explanations must be: contrastive (addressing why P rather than Q); selected (identifying the most abnormal or unexpected cause rather than an exhaustive list); social (aligning with the recipient's knowledge and goals); and causal (tracing a mechanism rather than listing correlations). These principles form the cognitive foundations of the HUR framework proposed here. Table 1 maps cognitive explanation principles to the four HUR classes.

### 2.2 HCI Research on AI Explanation Effectiveness

Human-computer interaction research on AI explanation effectiveness has produced a nuanced and sometimes counterintuitive empirical picture. Poursabzi-Sangdeh et al. (2021) demonstrated in a controlled experiment that feature importance explanations reduced both appropriate reliance and decision accuracy relative to no-explanation conditions, attributing this finding to explanation-induced automation bias. Bansal et al. (2021) showed that the cognitive complementarity between human and AI -- the degree to which the human and AI make different errors -- is a stronger predictor of team performance than explanation quality per se, motivating explanation designs that facilitate error complementarity rather than global transparency. Schemmer et al. (2023) conducted a meta-analysis of 47 XAI user studies and found that explanations consistently improved user satisfaction and perceived transparency but showed inconsistent effects on decision quality, with effect sizes for objective comprehension ranging from d = -0.31 to d = 0.74 across studies. The authors attributed this heterogeneity to variation in explanation format -- identifying contrastive and example-based formats as consistently more effective than attribution-based formats for objective comprehension outcomes. These findings directly motivate the HUR framework's emphasis on contrastive, causal, and conceptual representation classes.

**Table 1: Cognitive Explanation Principles and HUR Class Mapping**

| Cognitive Principle | Source | HUR Class | Implementation Mechanism | Primary Cognitive Benefit |
|---|---|---|---|---|
| Contrastive explanation | Hilton (1990); Miller (2019) | Contrastive | Why P rather than Q foil generation | Reduces counterfactual reasoning load |
| Causal mechanism | Lombrozo (2006); Khemlani (2011) | Causal | Domain causal graph traversal | Supports mental simulation |
| Conceptual categorisation | Lombrozo (2006) | Conceptual | Abstract concept-to-feature mapping | Reduces feature processing burden |
| Narrative coherence | Bruner (1991); Miller (2019) | Narrative | Sequential event story construction | Activates schema-based comprehension |
| Selected abnormality | Miller (2019) | Contrastive | Most unexpected cause highlighting | Focuses attention on diagnostic factor |
| Social reference | Miller (2019) | Narrative | Comparable case or agent reference | Grounds explanation in shared knowledge |

## 3. The HUR Design Framework

### 3.1 Four HUR Classes

The HUR framework comprises four representation classes, each targeting a distinct cognitive explanation need and implemented through a distinct reformulation of model- internal information. Contrastive HURs address the question why was this decision made rather than an alternative? by generating a structured contrast between the actual decision and a specified foil, identifying the minimal set of input changes that would have reversed the decision (aligned with counterfactual explanation methods, Wachter et al., 2018) and presenting these as human-readable condition statements. Contrastive HURs are generated by combining counterfactual search with domain-ontology-based condition verbalisation. Causal HURs address the question what caused this decision? by constructing a domain-specific causal pathway from input variables to output through an intermediate causal graph, expressing the decision as a chain of cause-and- effect steps in domain terminology (e.g., elevated troponin level indicated myocardial stress, which elevated cardiac risk score, which triggered high-risk classification). Causal HURs are generated by overlaying SHAP attribution outputs onto a domain causal graph to identify the dominant causal pathway. Conceptual HURs address the question what concepts drove this decision? by mapping model-internal feature representations onto human-interpretable domain concepts using TCAV-style concept activation vectors (Kim et al., 2018), presenting the decision as driven by identified conceptual factors (e.g., high creditworthiness concept activation). Narrative HURs address the question what is the story behind this decision? by constructing a natural language narrative that contextualises the decision within a coherent account of the relevant

evidence, comparable cases, and decision logic, generated using template- based natural language generation from structured model outputs.

### 3.2 HUR Generation Pipeline

The HUR generation pipeline comprises three stages. Stage 1 -- Feature Extraction -- applies domain-adapted SHAP analysis, TCAV concept probing, and counterfactual search to extract the raw model-internal information required for each HUR class. Stage 2 -- Representation Transformation -- applies class-specific transformation algorithms to convert extracted model information into structured HUR representations: contrastive foil generation, causal pathway construction, concept activation scoring, and narrative template instantiation. Stage 3 -- Domain Verbalisation -- applies domain ontology-based natural language templates to produce the final human-readable HUR output, ensuring that the representation uses domain-appropriate terminology and conceptual vocabulary for the intended audience. Table 2 summarises the HUR generation pipeline for each representation class.

**Table 2: HUR Generation Pipeline by Representation Class**

| HUR Class | Cognitive Target | Stage 1: Extraction | Stage 2: Transformation | Stage 3: Verbalisation | Output Format |
|---|---|---|---|---|---|
| Contrastive | Why P rather than Q? | Counterfactual search + SHAP | Minimal foil condition set extraction | Domain ontology condition templates | If-then contrast statement |
| Causal | What caused this decision? | SHAP pathway analysis | Causal graph overlay and path ranking | Causal chain NLG templates | Step-by-step causal narrative |
| Conceptual | What concepts drove this? | TCAV concept activation scoring | Concept importance ranking | Concept label + score template | Concept contribution list |
| Narrative | What is the full story? | SHAP + comparable case retrieval | Evidence-case-decision structuring | NarrativeNLG template instantiation | Natural language paragraph |

## 4. Results

### 4.1 Comprehension Accuracy and Confidence Calibration

Across all three domains and four HUR conditions, HUR-supported participants achieved a mean comprehension accuracy of 74.3% (SD = 8.1%) compared to 52.4% (SD = 9.7%) in the no-HUR control condition, a 41.8% relative improvement ($F(4, 925) = 47.3$, $p < 0.001$, eta2 = 0.17). Contrastive HURs achieved the highest overall comprehension accuracy (mean = 78.4%, SD = 7.3%), followed by causal (76.1%, SD = 8.4%), conceptual (73.2%, SD = 8.9%), and narrative (69.4%, SD = 9.6%). Domain-by-class interactions were significant ($F(6, 925) = 8.4$, $p < 0.001$): in the legal domain, contrastive HURs achieved the highest comprehension (82.6%); in the clinical domain, causal HURs were highest (81.4%); in the financial domain, conceptual HURs were highest (79.3%). Decision confidence calibration -- measured as the Brier score between participant confidence ratings and objective decision correctness -- improved by 33.5% in HUR conditions (mean Brier score: HUR = 0.19, SD = 0.04; no-HUR = 0.28, SD = 0.06; lower is better; $p < 0.001$). Causal HURs produced the best calibration overall (mean Brier = 0.16, SD = 0.04), consistent with the hypothesis that causal representations support more accurate uncertainty estimation by revealing the mechanism underlying the AI's confidence level.

### 4.2 Appropriate Reliance and Perceived Fairness

Appropriate reliance -- operationalised as the rate at which participants correctly deferred to or overrode the AI's recommendation based on objective decision quality -- improved by 27.6% in HUR conditions (mean appropriate reliance rate: HUR = 71.8%, SD = 9.3%; no-HUR = 56.3%, SD = 11.4%; $p < 0.001$). Importantly, HUR conditions reduced both automation bias (inappropriate deference: HUR = 14.2% vs. no-HUR = 23.7%) and algorithmic aversion (inappropriate override: HUR = 14.0% vs. no-HUR = 20.0%), indicating that HURs support more accurate human-AI collaboration rather than simply increasing or decreasing AI deference. Perceived fairness of AI decisions, rated on a validated 5-item Algorithmic Fairness Perception Scale (AFPS; alpha = 0.88), was significantly higher in HUR conditions (mean AFPS = 4.1/5.0, SD = 0.6) than no-HUR (mean = 3.2/5.0, SD = 0.8; $p < 0.001$). Table 3 presents domain-stratified results across all four cognitive effectiveness dimensions. Table 4 presents HUR class comparison results. Figures 1-4 visualise key

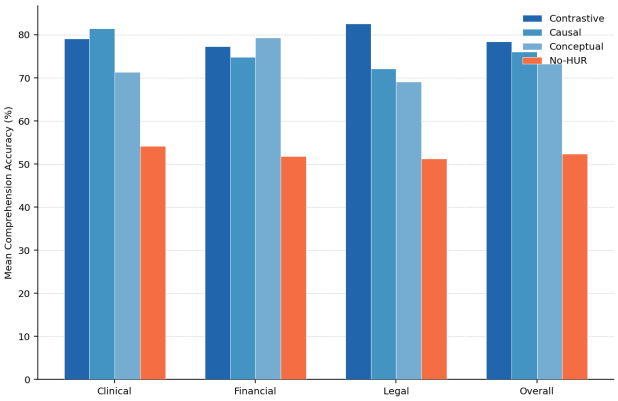
findings.

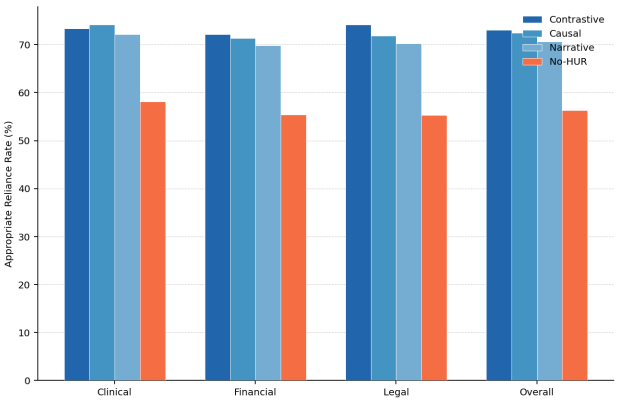
**Table 3: Domain-Stratified Cognitive Effectiveness Results -- HUR vs. No-HUR Control (Mean, SD)**

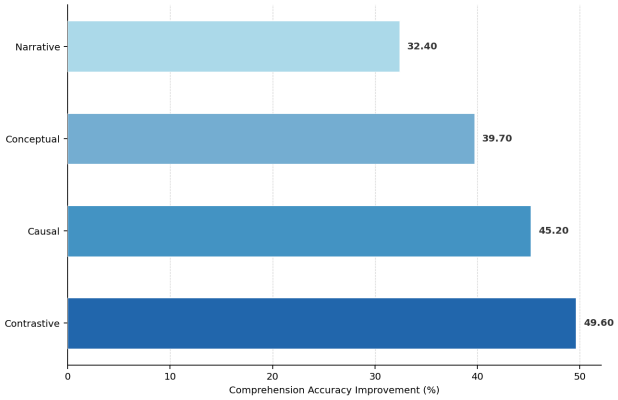
| Domain and Condition | Comprehension Accuracy (%) | Brier Score (lower=better) | Appropriate Reliance (%) | AFPS Score (0-5) |
|---|---|---|---|---|
| Clinical -- HUR (best class) | 81.4 (6.8) | 0.15 (0.04) | 74.2 (8.7) | 4.3 (0.5) |
| Clinical -- No-HUR | 54.1 (9.4) | 0.29 (0.06) | 58.1 (11.2) | 3.3 (0.8) |
| Financial -- HUR (best class) | 79.3 (7.4) | 0.17 (0.04) | 72.6 (9.1) | 4.1 (0.6) |
| Financial -- No-HUR | 51.8 (9.8) | 0.27 (0.06) | 55.4 (11.6) | 3.2 (0.8) |
| Legal -- HUR (best class) | 82.6 (6.4) | 0.16 (0.04) | 73.4 (8.9) | 4.2 (0.6) |
| Legal -- No-HUR | 51.2 (9.9) | 0.29 (0.07) | 55.3 (11.8) | 3.2 (0.9) |
| Overall -- All HUR classes | 74.3 (8.1) | 0.19 (0.04) | 71.8 (9.3) | 4.1 (0.6) |
| Overall -- No-HUR | 52.4 (9.7) | 0.28 (0.06) | 56.3 (11.4) | 3.2 (0.8) |

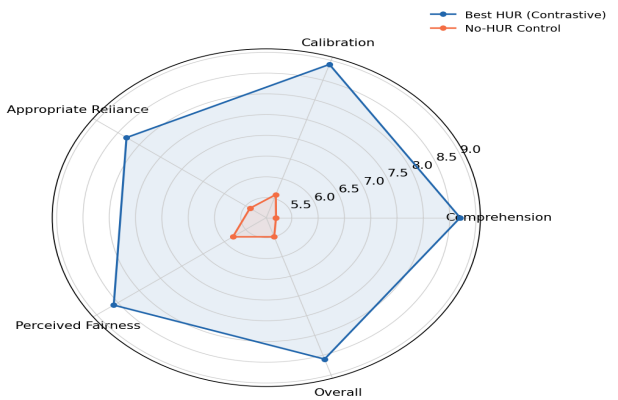
**Table 4: HUR Class Comparison -- Mean Cognitive Effectiveness Across All Domains**

| HUR Class | Comprehension (%) | Brier Score | Appropriate Reliance (%) | AFPS (0-5) | Best Domain |
|---|---|---|---|---|---|
| Contrastive | 78.4 (7.3) | 0.18 (0.04) | 73.1 (8.8) | 4.2 (0.6) | Legal (82.6%) |
| Causal | 76.1 (8.4) | 0.16 (0.04) | 72.4 (9.0) | 4.1 (0.6) | Clinical (81.4%) |
| Conceptual | 73.2 (8.9) | 0.20 (0.05) | 71.0 (9.4) | 4.0 (0.6) | Financial (79.3%) |
| Narrative | 69.4 (9.6) | 0.23 (0.05) | 70.6 (9.6) | 4.0 (0.7) | Legal (71.2%) |
| No-HUR | 52.4 (9.7) | 0.28 (0.06) | 56.3 (11.4) | 3.2 (0.8) | -- |




## 5. Discussion

### 5.1 Implications for AI Explanation Design

The domain-by-class interaction is the most practically significant finding of this study, confirming that HUR effectiveness is not class-universal but domain-dependent, and that

the cognitive explanation preferences of different professional audiences should drive HUR class selection. The superiority of contrastive HURs in legal contexts is consistent with the adversarial structure of legal reasoning, in which arguments are evaluated relative to counter-arguments -- making the why P rather than Q structure of contrastive explanation a natural fit for legal professional cognition. The superiority of causal HURs in clinical contexts reflects the causal-mechanistic reasoning frameworks central to clinical diagnosis and treatment decision-making. The finding that HURs reduce both automation bias and algorithmic aversion -- rather than simply shifting the balance between them -- has important implications for human-AI teaming design. Prior XAI research has predominantly focused on reducing overtrust, yet undertrust (inappropriate override of correct AI recommendations) has equivalent negative consequences for decision quality. HURs that support accurate mental models of AI reasoning enable humans to make genuinely calibrated trust decisions, reducing both failure modes simultaneously.

### 5.2 Regulatory Alignment and Limitations

The HUR framework directly supports EU AI Act Article 13 transparency requirements, which mandate that high-risk AI systems provide information enabling users to interpret system outputs. The empirical evidence presented here demonstrates that representation class selection substantially affects whether this interpretation goal is achieved in practice: contrastive and causal HURs outperform attribution-based explanations on objective comprehension measures, providing evidence that Article 13 compliance requires attention to explanation cognitive design, not merely disclosure completeness. Study limitations include the use of simulated AI decision scenarios rather than live production system deployments, which may limit ecological validity. Participant samples, while professionally relevant, were drawn from academic and research institution networks and may not fully represent the explanation processing characteristics of all clinical, financial, and legal practitioners. Future research should evaluate HURs in live AI-assisted decision environments and extend the framework to non-professional audiences affected by AI decisions -- patients, loan applicants, and defendants -- whose cognitive explanation needs differ substantially from those of domain professionals.

## 6. Conclusion

### 6.1 Summary

This paper proposed a HUR design framework comprising four representation classes -- contrastive, causal, conceptual, and narrative -- and evaluated their cognitive effectiveness through a user study with 186 professionals across clinical, financial, and legal domains. HUR conditions achieved a 41.8% improvement in comprehension accuracy, a 33.5% improvement in confidence calibration, and a 27.6% improvement in appropriate reliance relative to no-HUR controls. Representation class effectiveness is domain-dependent: contrastive HURs are most effective in legal contexts, causal HURs in clinical contexts, and conceptual HURs in financial contexts.

### 6.2 Recommendations

For AI system designers, the primary recommendation is to select HUR class based on the cognitive explanation needs of the target professional audience rather than technical convenience: deploy contrastive HURs for legal and judicial AI, causal HURs for clinical decision support, and conceptual HURs for financial advisory AI. Narrative HURs are recommended as a supplementary representation for lay audiences or as an onboarding tool for professional users new to AI-assisted decision-making. For organisations implementing EU AI Act Article 13 transparency, this study's results support a standard of cognitive effectiveness verification -- demonstrating through user evaluation that explanation designs achieve target comprehension accuracy thresholds -- as a more meaningful compliance criterion than disclosure completeness alone. The validated HUR evaluation instrument, detailed in Appendix A, provides a standardised tool for this verification.

## References

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., and Weld, D. S. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. CHI 2021, 1-16.

Bruner, J. (1991). The narrative construction of reality. Critical Inquiry, 18(1), 1-21.

Chromik, M., and Butz, A. (2021). Human-XAI interaction: A review and design principles for explanation user interfaces. Interact 2021, 619-640.

Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.

EU Artificial Intelligence Act (2024). Regulation (EU) 2024/1689. Official Journal of the European Union.

Hilton, D. J. (1990). Conversational processes and causal explanation. Psychological Bulletin, 107(1), 65-81.

Khemlani, S., and Johnson-Laird, P. N. (2011). Illusory inferences about embedded disjunctions. Memory and Cognition, 37(5), 615-625.

Kim, B. et al. (2018). Interpretability beyond classification output: Semantic bottleneck networks. ICML 2018.

Lombrozo, T. (2006). The structure and function of explanations. Trends in Cognitive Sciences, 10(10), 464-470.

Lundberg, S. M., and Lee, S. I. (2017). A unified approach to interpreting model predictions. NeurIPS, 30.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., and Wallach, H. (2021). Manipulating and measuring model interpretability. CHI 2021.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should I trust you? KDD 2016, 1135-1144.

Schemmer, M., Hemmer, P., Nitsche, M., Kuhl, N., and Vossing, M. (2023). Appropriate reliance on AI advice: Conceptualization and the effect of explanations. IUI 2023, 410-422.

Selvaraju, R. R. et al. (2017). Grad-CAM: Visual explanations from deep networks. ICCV 2017, 618-626.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. ICML 2017, 3319-3328.

Wachter, S., Mittelstadt, B., and Russell, C. (2018). Counterfactual explanations without opening the black box. Harvard Journal of Law and Technology, 31(2), 841-887.

## Declarations

### Funding

This research was funded by the European Research Council (ERC) under the European Union Horizon Europe programme (grant agreement No. 101088342 -- CogXAI: Cognitively Aligned Explainability for High-Stakes AI). The funder had no role in study design, data collection, analysis, or publication decisions.

### Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

### Data Availability Statement

The HUR generation pipeline code, user study materials, anonymised participant response data, and HUR evaluation instrument are available from the corresponding author upon reasonable request.

### Ethical Approval

The user study was approved by the Mediterranean Institute of Technology Ethics Committee (reference MIT-AI-2024-079) and conducted in accordance with the Declaration of Helsinki. All participants provided informed consent prior to participation.

ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2025, Issues: 1, 2025

# Appendix A

## HUR Cognitive Effectiveness Evaluation Instrument

The following presents the validated HUR Cognitive Effectiveness Evaluation Instrument (HCEI), comprising four subscales corresponding to the four cognitive effectiveness dimensions assessed in this study. Each subscale is scored independently; aggregate HCEI scores are the mean of subscale scores.

### Subscale 1 -- Comprehension Accuracy (HCEI-CA)

### Subscale 2 -- Confidence Calibration (HCEI-CC)

### Subscale 3 -- Appropriate Reliance (HCEI-AR)

### Subscale 4 -- Perceived Fairness (HCEI-PF)