Cognitive Decision Intelligence Framework for Explainable AI Systems

Artificial Intelligence (AI) is increasingly embedded in decision support systems, where its potential to enhance human reasoning is counterbalanced by risks of both overreliance and distrust. Explainable AI (XAI) has been introduced to mitigate these challenges, yet most approaches still overlook the complexity of human cognition, which cannot be reduced to purely formal processes. Cognitive psychology highlights multiple mechanisms, ranging from fast and intuitive to effortful and logical reactions, from mental model construction to bias awareness, that shape how people interpret explanations and integrate them into decision-making. To this end, we conducted a preregistered systematic review of experimental studies with human participants, focusing on XAI, and assessing which cognitive processes are studied and how. The review followed PRISMA guidelines, and out of 26,290 records, 229 full-texts were included. Alongside the Downs & Black methodological appraisal for *risk of bias*, we developed a novel *psychological risk of bias* tool to evaluate the conceptual precision and operationalization of how cognitive constructs are measured. Preliminary analyses indicate that XAI studies tend to cluster around four main groups of cognitive outcomes. The first concerns **trust and reliance**: reliance (appropriate or overreliance) is often assessed through behavioral and quantitative measures (e.g., weight of advice), while trust is usually measured via self-reports, with substantial variability and limited use of standardized scales or trust subfactors. The second cluster relates to **performance and decision quality**, including accuracy and reaction times. These measures are typically behavioral and standardized, yet a gap often emerges between the populations tested (frequently general users) and the domain expertise required in real-world decision contexts. The third cluster focuses on **information processing and workload**: XAI explanations can alter cognitive load, shape mental models, and influence attention and understanding. Finally, a fourth cluster addresses **attitudes and awareness**, including confidence, self-efficacy, and bias awareness. Taken together, these findings suggest that XAI influences several distinct yet interrelated cognitive domains. However, the current evidence remains fragmented, and it is not yet possible to determine whether XAI “works” in a general sense or to identify which types of explanations are effective across contexts. Different explanation strategies may affect different cognitive outcomes: for instance, counterfactual explanations may improve task performance but not necessarily foster trust. Overall, while XAI clearly impacts multiple cognitive domains, the systematic analysis reveals a major gap: the lack of systematic attention to psychological confounders, which are rarely investigated. Therefore, we designed a pilot experimental protocol that combines XAI with Learning to Defer to explore whether explanations and adaptive delegation can support analytical reasoning. A key focus is on the psychological profile of participants, specifically investigating *System 1 vs. System 2* activation. After each classification task, participants will self-assess their reasoning style using a pre-validated **Visual Assessment Scale**. In addition, dispositional traits will be measured, including **intolerance of uncertainty** (IUS-12), **attitudes toward AI** (General Attitudes towards Artificial Intelligence Scale), **emotion regulation strategies** (Emotion Regulation Questionnaire), and **general decision-making style** (GDMS). While exploratory, this pilot illustrates the kind of empirical validation needed to bridge theory and practice. Building on both the systematic synthesis and the experimental exploration, we propose a **Decision Intelligence Framework** that embeds psychological and ethical dimensions into XAI. A distinctive feature is the integration of psychological risk of bias assessment, extending methodological appraisal to ensure that explanations genuinely align with human cognition. The framework provides guidelines for designing explanations that balance intuitive and analytical reasoning, foster appropriate trust, and reduce cognitive biases. By combining large-scale evidence synthesis, exploratory experimentation, and theory-driven framework development, this work advances a new perspective on XAI, not only as a technical tool but as a cognitive aid. The expected outcome is a set of explainable, trustworthy, and psychologically grounded XAI systems that can sustain effective human-AI collaboration in complex and high-stakes decision-making.