Universidade de Pernambuco
Programa de Pós-Graduação em Engenharia da Computação (PPGEC)

Proposta de Tese de Doutorado

Área: Inteligência Computacional

Título: Confiança Apropriada e Contestabilidade em Decisões Assistidas por IA: Um Arcabouço Mensurável Aplicado à Gestão de Projetos de Software

Orientador: Profº Dr. Cleyton Mário de Oliveira Rodrigues
Coorientador: Profº Dr. Carlo Marcelo Revoredo da Silva

A adoção crescente de modelos de linguagem de grande porte (LLMs) em ambientes de desenvolvimento de software deslocou o principal gargalo da produção para a decisão sobre quando confiar nos artefatos gerados automaticamente. Evidências empíricas recentes revelam paradoxos relevantes: desenvolvedores experientes tornam-se menos produtivos ao usar assistentes de IA em repositórios maduros, embora percebam ganhos de velocidade; e combinações de humano e IA frequentemente decidem pior do que cada um isoladamente, com perdas concentradas justamente em tarefas de decisão. O problema central, portanto, não é a capacidade dos modelos, mas a calibração da confiança humana sobre suas saídas: o que a literatura denomina confiança apropriada.

Esta tese de doutorado propõe desenvolver e validar um arcabouço mensurável de confiança apropriada e contestabilidade para decisões assistidas por IA, com demonstração empírica no domínio da gestão de projetos de software. A pesquisa é estruturada em quatro questões centrais: como medir confiança apropriada e contestabilidade com validade de construto; quando a sugestão da IA melhora ou degrada a qualidade da decisão; quais intervenções que reduzem o custo de escrutínio melhoram causalmente a confiança apropriada; e se a contestabilidade por projeto melhora desfechos e calibração.

A metodologia adota Design Science Research (DSR), combinando construção e validação de instrumento de medida, experimentos pré-registrados com sujeitos humanos e implementação das intervenções em uma ferramenta real de gestão de projetos. O domínio de demonstração privilegia estimativas de tamanho funcional COSMIC, no qual o candidato já dispõe de motor simbólico auditável, que fornece proveniência verificável das decisões. As intervenções investigadas incluem explicação seletiva, abstenção, conjuntos de predição conformes, atrito deliberado em situações de alta incerteza e rastreabilidade de decisões automatizadas.

Espera-se contribuir com um conjunto validado de métricas, evidência causal sobre intervenções eficazes, um testbed com bases abertas para replicação e mapeamento dos instrumentos aos requisitos de supervisão do Regulamento Europeu de IA. A pesquisa posiciona-se em um problema reconhecidamente aberto na interseção entre engenharia de software, interação humano-computador e IA confiável.

Referências Bibliográficas:

- ALFRINK, K. et al. Contestable AI by design: towards a framework. *Minds and Machines*, 2022.
- ANGELOPOULOS, A. N.; BATES, S. Conformal prediction: a gentle introduction. *Foundations and Trends in Machine Learning*, 2023.
- BALTES, S. et al. Guidelines for empirical studies in software engineering involving large language models. *Empirical Software Engineering*, 2025.
- BANSAL, G. et al. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In: *CHI Conference on Human Factors in Computing Systems*, 2021.

- BUÇINCA, Z.; MALAYA, M. B.; GAJOS, K. Z. 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 (CSCW)*, 2021.
- DVIJOTHAM, K. et al. Enhancing the reliability and accuracy of AI-enabled diagnosis via complementarity-driven deferral to clinicians (CoDoC). *Nature Medicine*, 2023.
- FOK, R.; WELD, D. S. In search of verifiability: explanations rarely enable complementary performance in AI-advised decision making. *AI Magazine*, 2024.
- GOH, E. et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Network Open*, 2024.
- HEVNER, A. R. et al. Design science in information systems research. *MIS Quarterly*, v. 28, n. 1, p. 75–105, 2004.
- METR (BECKER, J. et al.). Measuring the impact of early-2025 AI on experienced open-source developer productivity. arXiv:2507.09089, 2025.
- MILLER, T. Explainable AI is dead, long live explainable AI! Hypothesis-driven decision support using evaluative AI. In: *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2023.
- MOZANNAR, H.; SONTAG, D. Consistent estimators for learning to defer to an expert. In: *International Conference on Machine Learning (ICML)*, 2020.
- PERRY, N. et al. Do users write more insecure code with AI assistants? In: *ACM Conference on Computer and Communications Security (CCS)*, 2023.
- TURPIN, M. et al. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. In: *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- UNIÃO EUROPEIA. *Regulamento (UE) 2024/1689 do Parlamento Europeu e do Conselho* (AI Act), Anexo III e Artigos 12, 14, 26 e 86, 2024.
- VACCARO, M.; ALMAATOUQ, A.; MALONE, T. When combinations of humans and AI are useful: a systematic review and meta-analysis. *Nature Human Behaviour*, 2024.
- VASCONCELOS, H. et al. Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, 2023.