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

Proposta de Dissertação de Mestrado

Área: Computação Inteligente ou Modelagem Computacional

Título: Criando e Avaliando Treinadores de IA Confiáveis para Jogos Competitivos

Orientador(a): Pablo Barros (pvab@ecom.poli.br)
Co-Orientador(a): Ana Tanevska (ana.tanevska@it.uu.se) (Uppsala University, Sweden)

Descrição:

Este projeto de pesquisa explora o desenvolvimento e a implementação de um agente virtual baseado em diálogo projetado para aprimorar a confiança e o engajamento por meio de técnicas de IA explicável (XAI). O foco está em avançar a interação humano-computador, permitindo que agentes artificiais forneçam explicações transparentes e compreensíveis para seus processos de tomada de decisão.

A confiança em agentes artificiais é uma área crítica de pesquisa [1], mas os sistemas de aconselhamento existentes frequentemente falham em oferecer respostas empáticas e compreensíveis, limitando sua capacidade de gerar confiança e interação prolongada [2]. Uma direção para abordar essa questão é capacitar o agente artificial a ser transparente sobre seu raciocínio, utilizando mecanismos de explicabilidade que permitam fornecer informações sobre seu funcionamento interno, de forma que o parceiro humano possa inferir como e por que o agente se comporta da maneira que se comporta [3].

A solução proposta envolve a criação de um agente consultivo específico para jogos, com o objetivo de ensinar e orientar usuários a jogar o jogo de cartas **Chef's Hat**. Serão desenvolvidas duas versões do agente para análise comparativa:

1. **Agente Consultivo Explicável:** Incorpora frameworks de XAI (por exemplo, SHAP ou LIME) para gerar explicações em linguagem natural sobre decisões no jogo. Essas explicações serão projetadas para serem acessíveis e compreensíveis (por exemplo: "Escolhi esta carta porque ela aumenta suas chances de vencer, maximizando suas combinações de alto valor").
2. **Agente Não Explicável:** Fornece respostas diretas, sem oferecer explicações sobre suas decisões.

A metodologia de avaliação envolve a realização de estudos de interação humano-agente (HAI) para analisar as duas versões do agente. Os participantes jogarão **Chef's Hat** com ambos os agentes, e suas experiências serão medidas em três dimensões:

- **Confiança:** Avaliada usando a **Escala de Confiança em Automação** (TIAS), para medir a confiança dos participantes na orientação do agente.
- **Engajamento:** Avaliado por meio de métricas observacionais, como frequência de interação e disposição para continuar jogando.
- **Aprendizado:** Medido pelo desempenho dos participantes no jogo e suas autoavaliações sobre a compreensão das mecânicas do jogo.

Os estudos planejados também utilizarão diretrizes existentes para projetar IA/HAI confiáveis (como o **ALTAI [4]**), para identificar quais aspectos da interação

(transparência do agente, senso de agência do usuário, etc.) mais contribuem para a confiança e o engajamento dos usuários em relação ao agente consultivo.

Este projeto contribui para o avanço no design de sistemas de IA confiáveis ao combinar explicabilidade com estratégias eficazes de interação humano-agente. O resultado esperado é uma compreensão mais profunda de como as técnicas de XAI podem ser utilizadas para construir confiança e engajamento, promovendo uma experiência do usuário mais satisfatória e significativa em cenários de aconselhamento.

Este projeto será conduzido em um ambiente de colaboração internacional, exigindo um domínio avançado do inglês para comunicação eficaz.

Referências Bibliográficas:

[1] Siau, K., & Wang, W. (2018). *Building trust in artificial intelligence, machine learning, and robotics*. ACM Transactions on Management Information Systems, 9(2), 1-37.
[2] Lombrozo, T. (2006). *The structure and function of explanations*. Trends in Cognitive Sciences, 10(10), 464–470.
[3] Wallkötter, S., Tulli, S., Castellano, G., Paiva, A., & Chetouani, M. (2021). Explainable embodied agents through social cues: a review. ACM Transactions on Human-Robot Interaction (THRI), 10(3), 1-24.
[4] Ala-Pietilä, Pekka, Yann Bonnet, Urs Bergmann, Maria Bielikova, Cecilia Bonefeld-Dahl, Wilhelm Bauer, Loubna Bouarfa et al. The assessment list for trustworthy artificial intelligence (ALTAI). European Commission, 2020.