Abrir o jogo compensa? Modelagem computacional da delação premiada via Teoria dos Jogos com agentes baseados em modelos de linguagem

Joao Victor Silva de Sousa [1]
Departamento Acadêmico de Educação Básica e Formação de Professores
Instituto Federal do Amazonas
Manaus, Brasil

Tacildo de Souza Araújo [2]
Departamento Acadêmico de Educação Básica e Formação de Professores
Instituto Federal do Amazonas
Manaus, Brasil

14 de janeiro de 2026

Palavras-chave: Teoria dos Jogos; Dilema do Prisioneiro; Modelagem Computacional; Modelos de Linguagem; Delação Premiada.

Resumo

Este trabalho propõe uma modelagem computacional da delação premiada a partir da Teoria dos Jogos, interpretando esse instituto jurídico como uma instância do Dilema do Prisioneiro em jogos repetidos. A delação envolve uma decisão estratégica sob incerteza, na qual agentes avaliam riscos e benefícios associados à cooperação ou à traição. Tal estrutura é classicamente associada a jogos não cooperativos de soma não zero, amplamente discutidos na literatura de Teoria dos Jogos (Fiani, 2009; Axelrod, 1984). A questão central investigada é se agentes baseados em modelos de linguagem reproduzem, desviam ou subvertem os padrões estratégicos previstos pela Teoria dos Jogos clássica.

A partir dessa formalização, investigamos o comportamento de agentes baseados em Modelos de Linguagem (Language Models – LMs), implementados em ambiente Python e tratados como agentes heurísticos capazes de tomar decisões estratégicas sem pressupor racionalidade clássica. Diferentemente de abordagens tradicionais, não são impostas estratégias racionais pré-definidas, permitindo que as decisões emerjam a partir do contexto e do histórico das interações (Russell & Norvig, 2013).

A metodologia consiste na definição de uma matriz de payoff associada à delação premiada, com estratégias binárias de cooperação e traição, seguida da simulação iterativa de interações estratégicas em jogos repetidos. As simulações foram conduzidas ao longo de 200 rodadas, considerando um agente heurístico inspirado em modelos de linguagem e uma estratégia clássica do tipo Tit-for-Tat. Os resultados foram analisados por meio de métricas quantitativas como taxa de cooperação e payoff médio por interação. O equilíbrio de Nash é apresentado como

[1]e-mail: 2024007388@ifam.edu.br
[2]e-mail: tacildo.araujo@ifam.edu.br

referência teórica no caso estático do Dilema do Prisioneiro (Nash, 1950), sendo utilizado como contraponto analítico, e não como objetivo normativo, na interpretação dos comportamentos observados.

Os experimentos indicaram a emergência de cooperação estável ao longo das interações, com o agente baseado em modelo de linguagem apresentando uma taxa de cooperação de aproximadamente 100% das rodadas simuladas. O payoff médio observado foi de 3,0 por interação para ambos os agentes, valor compatível com o resultado de cooperação mútua no Dilema do Prisioneiro. Esses resultados evidenciam comportamentos estratégicos fora do equilíbrio clássico de Nash, especialmente em jogos repetidos, sugerindo que agentes heurísticos baseados em linguagem podem sustentar padrões cooperativos estáveis.

Como limitação do modelo, destaca-se que a abordagem proposta abstrai aspectos jurídicos complexos da delação premiada, como a exigência de provas materiais para a validação das informações fornecidas. Tal simplificação é intencional e visa isolar a dimensão estratégica da decisão de cooperar ou não, sem a pretensão de reproduzir integralmente a complexidade normativa do instituto. Ainda assim, os resultados obtidos evidenciam o potencial da modelagem matemática e da simulação computacional como ferramentas da Matemática Aplicada para a análise de sistemas estratégicos associados a fenômenos sociais e jurídicos contemporâneos.

Referências

[1] D. G. de Figueiredo, *O Princípio de Dirichlet*, Revista Matemática Universitária, 1985.