Jogos de Sinalização para Agentes de IA Comprometidos

Lucila M. S. Bento[1], **Davidson R. Boccardo**[2]

[1]Depto. de Informática e Ciência da Computação – Instituto de Matemática e Estatística
Universidade do Estado do Rio de Janeiro (UERJ)

[2]Einstein Hospital Israelita

lucila.bento@ime.uerj.br, davidson.boccardo@einstein.br

Abstract. *AI agents may hold valid delegated credentials while operating under compromised context. Adaptive authorization is modeled as a Bayesian signaling game in which an intact or compromised agent emits a basic or reinforced containment signal, a detector produces noisy evidence, and a verifier grants, challenges, or denies access. Decision thresholds and separating conditions are derived, binary and graduated authorization are compared, and D1 refinement is implemented through linear programming over contingent mixed responses. In a synthetic parameter-grid evaluation, graduated authorization provides strictly higher utility in a region of posterior beliefs for 38.2% of the evaluated configurations. Stronger containment can nevertheless stabilize pooling, showing that improved defensive utility does not imply type separation.*

Resumo. *Agentes de IA podem possuir credenciais delegadas válidas e operar sob um contexto comprometido. Neste trabalho, a autorização adaptativa é modelada como um jogo de sinalização Bayesiano, no qual um agente íntegro ou comprometido emite um sinal básico ou reforçado, um detector produz evidência ruidosa e o verificador concede, desafia ou nega acesso. São derivados limiares e condições de separação, comparadas políticas binária e graduada e aplicado o refinamento D1 por programação linear. Em uma avaliação computacional com parâmetros sintéticos, a política graduada apresenta utilidade estritamente maior em uma região de crenças posteriores para 38,2% das configurações avaliadas. Entretanto, contenções mais fortes podem estabilizar pooling, mostrando que maior utilidade defensiva não implica separação dos tipos.*[1]

1. Introdução

Os agentes de IA baseados em grandes modelos de linguagem combinam mecanismos de memória e de recuperação de dados com o acesso a ferramentas, o que amplia a superfície de ataque. Greshake et al. [Greshake et al. 2023] mostram que instruções maliciosas inseridas em dados externos podem influenciar aplicações integradas a LLMs sem interação direta do atacante. O *AgentDojo* avalia esse tipo de *prompt injection* indireto em agentes com ferramentas [Debenedetti et al.], enquanto o OWASP GenAI Security Project classifica *prompt injection* entre os principais riscos de aplicações baseadas em LLMs e recomenda medidas como mínimo privilégio e aprovação humana para operações de

[1]Ferramentas de IA generativa (ChatGPT e Claude) foram usadas na revisão de linguagem e na análise de cobertura do código-fonte, verificando a consistência da implementação em relação ao modelo proposto.

maior risco [OWASP Foundation 2025]. O NIST também sistematiza ataques e medidas de mitigação ao longo do ciclo de vida de sistemas de IA [Vassilev et al. 2025].

Nesse cenário, autenticar corretamente um agente e verificar que uma operação é permitida pela política aplicável não garantem sua integridade operacional. Se a autoridade decorre de delegação, a cadeia e o escopo delegados também podem ser válidos e, ainda assim, o agente estar sob influência de conteúdo não confiável, memória contaminada, ferramentas comprometidas ou outros agentes. O problema considerado neste artigo começa precisamente nesse ponto: a autoridade formal para executar uma operação já foi estabelecida, mas permanece incerteza sobre o estado operacional do agente.

Este artigo propõe uma camada complementar de autorização adaptativa que não substitui autenticação nem a avaliação convencional de políticas. A interação posterior a essas verificações é modelada como um jogo de sinalização com evidência, no qual o agente escolhe entre um modo básico e um compromisso reforçado de contenção, um detector externo fornece evidência ruidosa e o decisor adaptativo mantém a concessão, impõe um desafio/contenção ou nega a operação. A contribuição combina autoridade formal válida, integridade operacional privada, contenção verificável, evidência imperfeita e uma decisão graduada. Desta forma, as principais contribuições são: **(i)** um modelo de ameaça e um jogo de sinalização para agentes formalmente autorizados, mas potencialmente comprometidos; **(ii)** condições analíticas para a política graduada e para a compatibilidade de incentivos de um equilíbrio separador; **(iii)** comparação entre autorizações binárias e graduadas; e **(iv)** um artefato reprodutível com enumeração de equilíbrios Bayesianos perfeitos (*Perfect Bayesian Equilibria*, PBE), busca numérica, refinamento D1 por programação linear, testes automatizados e mapas de parâmetros.

2. Trabalhos Relacionados

Trabalhos recentes fortalecem diferentes componentes da identidade e da autoridade de agentes. Pappu et al. [Pappu et al. 2025] usam SPIFFE/SPIRE para identidade de carga de trabalho e autenticação *zero trust*. Singh et al. [Singh et al. 2025] analisam registros para descoberta e identidade verificável, enquanto Fujie et al. [Fujie et al. 2026] distinguem identidade, verificação e autorização em espaços de dados orientados a agentes. O Lineage protege cadeias de delegação em múltiplos saltos [Stefanescu and Aciobăniței 2026], e Kolhe et al. [Kolhe et al. 2026] aplicam políticas externas às ferramentas e aos dados acessados via *Model Context Protocol* (MCP). Esses mecanismos estabelecem quem atua e qual autoridade pode exercer. O problema considerado neste trabalho começa depois dessas verificações, quando a integridade da execução permanece incerta.

Essa incerteza é particularmente relevante sob *prompt injection*. Greshake et al. [Greshake et al. 2023] mostram que instruções em dados recuperados podem controlar aplicações com LLMs. O *AgentDojo* avalia ataques indiretos contra agentes [Debenedetti et al.], e o *Task Shield* verifica alinhamento entre instruções, ações e objetivos [Jia et al. 2025]. O OWASP recomenda reduzir privilégios e exigir aprovação para ações críticas [OWASP Foundation 2025]. Este trabalho não propõe um novo detector, mas sim investiga como evidências externas de comprometimento e mecanismos de contenção podem alterar uma decisão já autorizada pela política convencional.

A literatura sobre jogos de sinalização e confiança fornece a base formal. Lu et al. [Lu et al. 2015] relacionam autorização a níveis de confiança, Pawlick et

al. [Pawlick et al. 2019] incorporam evidência probabilística a jogos de sinalização, e Boudagdigue et al. [Boudagdigue et al. 2023] combinam certificados, confiança e atualização Bayesiana em IoT. Esposito et al. [Esposito et al. 2020] são particularmente próximos por considerarem nós com credenciais legítimas que passam a apresentar comportamento malicioso. A diferença deste trabalho é combinar, para agentes de IA, uma operação já formalmente autorizada, compromisso verificável de contenção, evidência ruidosa e três respostas possíveis. Para múltiplos equilíbrios, o critério D1 é aplicado [Banks and Sobel 1987], relacionado ao refinamento de Cho e Kreps [Cho and Kreps 1987], sobre conjuntos completos de respostas mistas.

3. Modelo de Ameaça e Jogo

Os ativos protegidos incluem ferramentas, APIs, dados, memória e efeitos externos acessíveis ao agente. O adversário busca induzir um agente legítimo a agir contra a intenção do usuário ou delegante. O comprometimento pode decorrer de *prompt injection* indireta, conteúdo malicioso recuperado por *Retrieval-Augmented Generation* (RAG), memória contaminada, ferramentas comprometidas ou comunicação entre agentes. O modelo não pressupõe falsificação de identidade, controle do decisor ou conhecimento perfeito do detector.

Antes do jogo, a identidade do agente já foi autenticada e a política convencional de autorização já permite a operação solicitada. Quando a autoridade decorre de delegação, sua validade e escopo também são verificados; delegação não é, portanto, requisito universal de controle de acesso. O jogo atua após essas checagens, quando a autoridade é válida, mas há incerteza operacional. Essa etapa usa evidências externas de monitoramento, não do controle de acesso, para que o decisor adaptativo mantenha, contenha ou bloqueie a operação — sem exigir inspeção de todos os recursos, nem que a autorização detecte o comprometimento diretamente.

A Figura 1 resume essa sequência: após as verificações convencionais, o agente escolhe um modo de execução, um detector externo produz evidência sobre possível comprometimento e o decisor adaptativo escolhe entre manter a concessão, impor contenção ou negar a operação.

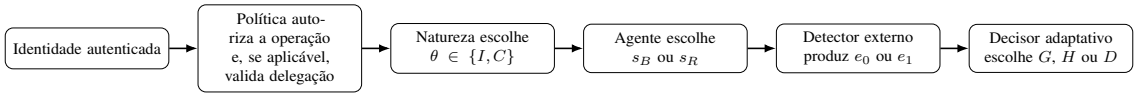

Figura 1. Fluxo da autorização convencional à decisão adaptativa.

Dada uma operação previamente autorizada, a Natureza escolhe o tipo $\theta \in \{I, C\}$, íntegro ou comprometido, com probabilidade a priori $p = \Pr(C)$. O agente observa seu tipo e escolhe $s \in \{s_B, s_R\}$. Em s_B, a execução mantém as condições usuais. Em s_R, o agente aceita executar sob um perfil reforçado oferecido e imposto ou atestado pela infraestrutura, por exemplo, com menor escopo, isolamento de memória, remoção de ferramentas ou aprovação humana.

O comprometido não é impedido de escolher s_R: justamente por isso a possibilidade de imitação faz parte do jogo. O custo c_C não precisa representar maior custo computacional de ativar a contenção; ele também captura a perda estratégica sofrida pelo

comprometido quando restrições eliminam ações ou ferramentas úteis ao ataque. O termo $\rho_R G_C$ representa o benefício ofensivo ainda possível dentro do escopo residual. O modelo não pressupõe estruturalmente que $c_C > c_I$. O detector externo produz $e \in \{e_0, e_1\}$, ausência ou presença de alerta. Para cada sinal s, $\alpha_s = \Pr(e_1 \mid I, s)$ e $\beta_s = \Pr(e_1 \mid C, s)$, com $\beta_s > \alpha_s$. Se $\sigma_I(s)$ e $\sigma_C(s)$ são as probabilidades de emissão, a crença posterior de comprometimento é

$$\mu(C \mid s, e) = \frac{p\,\sigma_C(s) P(e \mid C, s)}{p\,\sigma_C(s) P(e \mid C, s) + (1-p)\sigma_I(s) P(e \mid I, s)}. \tag{1}$$

Com base em (s, e), o decisor escolhe $a \in \{G, H, D\}$: manter a concessão, desafiar/conter ou negar.

3.1. Utilidades

Ao manter a concessão, o decisor recebe benefício B_V se o agente for íntegro e sofre perda L_C se estiver comprometido. Negar uma operação legítima gera perda L_I. A ação H custa C_H, causa perda L_H^I ao íntegro e limita a perda a L_H^C sob comprometimento. Para $\mu = \Pr(C \mid s, e)$, suas utilidades esperadas são $U_V(G \mid \mu) = (1-\mu)B_V - \mu L_C$, $U_V(H \mid \mu) = (1-\mu)(B_V - C_H - L_H^I) - \mu(C_H + L_H^C)$ e $U_V(D \mid \mu) = -(1-\mu)L_I$.

Para o íntegro, B_I é o benefício da operação, $c_I(s)$ o custo do modo escolhido, h_I o custo do desafio e d_I a perda por negação: $U_I(s, G) = B_I - c_I(s)$, $U_I(s, H) = B_I - h_I - c_I(s)$ e $U_I(s, D) = -d_I - c_I(s)$. Para o comprometido, G_C é o ganho potencial e ρ_s e ρ_{Hs} são frações residuais desse ganho sob concessão e desafio. O parâmetro $c_C(s)$ representa o custo do modo de execução, h_C o custo do desafio e r_C o custo adicional imposto quando a tentativa é contida ou bloqueada: $U_C(s, G) = \rho_s G_C - c_C(s)$, $U_C(s, H) = \rho_{Hs} G_C - h_C - r_C - c_C(s)$ e $U_C(s, D) = -r_C - c_C(s)$.

O sinal básico satisfaz $c_I(s_B) = c_C(s_B) = 0$ e $\rho_{s_B} = 1$. Para s_R, $c_I(s_R) = c_I$, $c_C(s_R) = c_C$ e $\rho_{s_R} = \rho_R$; sob desafio, as frações residuais são ρ_{HB} e ρ_{HR}.

Os valores da Tabela 1 definem uma configuração-base sintética, construída para representar relações qualitativas coerentes com o cenário modelado, e não estimativas obtidas de um sistema real. Os parâmetros foram normalizados em torno de benefícios unitários e escolhidos de modo que o comprometimento seja minoritário, o detector seja informativo ($\beta_s > \alpha_s$), uma concessão indevida seja mais onerosa que uma negação indevida ($L_C > L_I$) e a contenção reduza, sem necessariamente eliminar, o impacto de um agente comprometido ($L_H^C < L_C$). Assim, $p = 0{,}20$ representa um cenário de comprometimento minoritário, mas não desprezível. Os demais valores instanciam custos e perdas relativos compatíveis com essas premissas. Como não foram calibrados empiricamente, a sensibilidade dos resultados à variação dos parâmetros centrais é avaliada posteriormente.

4. Análise de Equilíbrio

Como as utilidades do decisor são afins em μ, os pontos de mudança entre ações são obtidos diretamente pelas condições de indiferença $U_V(G) = U_V(H)$, $U_V(H) = U_V(D)$ e $U_V(G) = U_V(D)$: $\tau_{GH} = \frac{C_H + L_H^I}{L_C + L_H^I - L_H^C}$, $\tau_{HD} = \frac{B_V + L_I - C_H - L_H^I}{B_V + L_I - L_H^I + L_H^C}$, e $\tau_{GD} = \frac{B_V + L_I}{B_V + L_I + L_C}$.

Tabela 1. Parâmetros da configuração-base.

Símbolo	Significado	Valor
p	probabilidade de comprometimento	0,20
(α_B, β_B)	detector sob s_B	(0,15; 0,70)
(α_R, β_R)	detector sob s_R	(0,05; 0,90)
(B_V, L_C, L_I)	benefício e perdas do decisor	(1; 4; 1)
(C_H, L_H^I, L_H^C)	custo e perdas de H	(0,20; 0,15; 0,50)
(B_I, G_C)	benefícios dos agentes	(1; 1)
(c_I, h_I, d_I)	custos do íntegro	(0,10; 0,15; 0,75)
(h_C, r_C)	custos do comprometido	(0,10; 0,20)
(ρ_{HB}, ρ_{HR})	ganho residual sob H	(0,30; 0)
(ρ_R, c_C)	contenção e custo de imitação	$[0,1]^2$

Quando os cruzamentos estão ordenados como $0 < \tau_{GH} < \tau_{GD} < \tau_{HD} < 1$, como na configuração-base, o decisor mantém a concessão para baixa probabilidade posterior de comprometimento, aplica H na região intermediária e nega na região alta. Na configuração-base, $\tau_{GH} \approx 0{,}096, \tau_{GD} \approx 0{,}333, \tau_{HD} \approx 0{,}702$. Uma vez estimados os parâmetros, esses limiares são constantes e podem ser pré-calculados. O modelo não requer reavaliar o estado de todos os recursos a cada solicitação.

Proposição 1 (condição de compatibilidade de incentivos). Considere o perfil separador $I \to s_R$ e $C \to s_B$. Nesse perfil, Bayes implica $\mu(C \mid s_R, e) = 0$ e $\mu(C \mid s_B, e) = 1$ para evidências no caminho de equilíbrio. Na configuração considerada, G é a melhor resposta para $\mu = 0$ e D para $\mu = 1$. Dados esses comportamentos sequencialmente racionais do decisor, nenhum tipo possui desvio lucrativo se, e somente se,

$$c_I \leq B_I + d_I \qquad \text{e} \qquad c_C \geq \rho_R G_C + r_C. \tag{2}$$

Demonstração. O agente íntegro recebe $B_I - c_I$ ao escolher s_R e $-d_I$ ao desviar para s_B. Logo, não desvia se $B_I - c_I \geq -d_I$, isto é, $c_I \leq B_I + d_I$. O comprometido recebe $-r_C$ ao permanecer em s_B e $\rho_R G_C - c_C$ ao imitar s_R e obter concessão. Portanto, não desvia se $-r_C \geq \rho_R G_C - c_C$, equivalente a $c_C \geq \rho_R G_C + r_C$. □

A segunda condição da Proposição 1 pode ser escrita como $\rho_R \leq (c_C - r_C)/G_C$. Assim, a separação pode ser favorecida tanto pela redução do ganho residual disponível ao comprometido quanto pelo aumento do custo associado à imitação do modo reforçado. A Figura 2 ilustra essa relação para a configuração-base: a região acima da fronteira satisfaz a condição de compatibilidade de incentivos do agente comprometido para o perfil separador, enquanto abaixo dela a imitação de s_R pode ser vantajosa.

Quando existem múltiplos equilíbrios, é aplicado o critério D1. Para um pooling em s e um desvio para s', seja q um plano contingente misto do decisor. Seja $\Delta_\theta(q) = V_\theta(s', q) - U_\theta^*$ e $W_\theta = \{q : \Delta_\theta(q) \geq 0\}$. O tipo θ é eliminado em favor de θ' quando $W_\theta \subset \{q : \Delta_{\theta'}(q) > 0\}$.

A inclusão do D1 é verificada por programação linear, minimizando $\Delta_{\theta'}(q)$ sujeito a $\Delta_\theta(q) \geq 0$ e às restrições de simplex. Como os ganhos são lineares em q, um mínimo estritamente positivo estabelece a inclusão para todas as respostas mistas admissíveis.

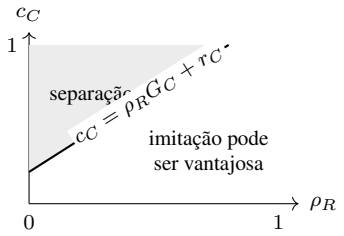

Figura 2. Fronteira de compatibilidade de incentivos para $G_C = 1$ e $r_C = 0{,}20$.

No pooling reforçado da configuração-base, uma derivação disponível no artefato fornece $V_C^B(q) \geq \frac{7}{12}(V_I^B(q) - 0{,}85)$. Como $V_I^B(q) \geq U_I^R$ implica $V_C^B(q) > U_C^R$ em $(\rho_R, c_C) \in [0,1]^2$, o D1 atribui o desvio ao comprometido. Portanto, todo pooling reforçado existente nesse domínio sobrevive ao refinamento, independentemente da discretização.

5. Avaliação Computacional

A avaliação computacional utiliza uma implementação em Python, disponível em `https://github.com/haTfk7JKY69cVADy9/GameAgent`, que enumera PBE puros, busca equilíbrios mistos, verifica arrependimentos e aplica D1 por programação linear. O artefato reúne código, parâmetros, sementes, resultados em CSV, scripts de reprodução e 32 testes automatizados. São avaliadas as perguntas: **RQ1**, quando a separação é sustentável; **RQ2**, quando a evidência altera a decisão; **RQ3**, quando H melhora a política binária; e **RQ4**, quais *poolings* sobrevivem ao D1.

Como instanciação ilustrativa, considera-se um agente já autenticado e autorizado a usar ferramentas via MCP. Em s_R, ferramentas de alto risco são removidas e ações externas exigem aprovação. Com os parâmetros-base, sob pooling básico, $\mu(C \mid e_0) \approx 0{,}081$ leva a G e $\mu(C \mid e_1) \approx 0{,}538$ a H; sob pooling reforçado, os valores correspondentes são $0{,}026$ e $0{,}818$, levando a G e D. O exemplo apenas ilustra a parametrização e não constitui validação empírica.

Para a RQ1, ρ_R e c_C variam em uma grade 21×21. A condição da Equação 2 coincide com a classificação computacional nas 441 configurações, sem falsos positivos ou negativos. Os 153 pontos separadores são exatamente os que satisfazem $\rho_R \leq (c_C - r_C)/G_C$, fornecendo uma verificação cruzada entre a derivação analítica e sua implementação.

Para a RQ2, sob pooling básico, são usados detectores sintéticos fraco, moderado e forte, com $(\alpha, \beta) = (0{,}25, 0{,}55)$, $(0{,}10, 0{,}75)$ e $(0{,}03, 0{,}95)$, respectivamente. Eles produziram ações distintas após e_0 e e_1 em 38,5%, 88,2% e 97,6% dos 999 valores de $p \in [0{,}001, 0{,}999]$. Portanto, maior capacidade discriminativa do detector só altera a ação quando a evidência desloca a crença posterior atravessa um limiar decisório. Esses perfis não foram estimados a partir do AgentDojo ou do Task Shield.

Para a RQ3, a comparação considera $\{G, D\}$ e $\{G, H, D\}$ em 9.261 combinações sintéticas de C_H, L_H^I e L_H^C, avaliando cada uma em 1.001 valores de $\mu \in [0, 1]$. Em 38,2% das configurações existe uma região de crenças em que a política graduada supera estritamente a binária. Sob distribuição uniforme de μ, o ganho esperado médio foi 0,033 e o máximo 0,667. Esses números caracterizam a grade sintética e não correspondem a ganhos observados em sistemas reais. Repetindo a grade para $p \in \{0{,}05, 0{,}20, 0{,}50\}$, $L_C \in \{2, 4, 8\}$ e $r_C \in \{0{,}10, 0{,}20, 0{,}40\}$, o número de pontos separadores foi 190, 153 e

91 à medida que r_C aumentou, como previsto pela fronteira analítica. A sobrevivência do pooling básico variou de 1 a 336 configurações, enquanto o pooling reforçado sobreviveu ao D1 sempre que existiu.

Assim, os percentuais dependem da parametrização, mas a fronteira analítica e a robustez do pooling reforçado se mantiveram. Para a RQ4, a Tabela 2 resume os equilíbrios de pooling existentes e sobreviventes ao refinamento: o pooling básico sobreviveu em 153 das 441 configurações sob a política binária, 120 sob a graduada-base e 210 no desafio eficaz e barato, enquanto o pooling reforçado sobreviveu sempre que constituía um PBE.

Tabela 2. Sobrevivência dos equilíbrios de pooling após o D1.

Política	s_B existe	s_B sobrevive	s_R existe	s_R sobrevive
Binária	441	153 (34,7%)	41	41 (100%)
Graduada-base	441	120 (27,2%)	41	41 (100%)
Desafio eficaz	441	210 (47,6%)	7	7 (100%)

6. Discussão, Limitações e Conclusão

Os resultados estendem a literatura ao combinar uma operação formalmente autorizada com incerteza sobre a integridade operacional. Lu et al. [Lu et al. 2015] relacionam autorização a confiança, enquanto Pawlick et al. [Pawlick et al. 2019] analisam evidência probabilística em jogos de sinalização. Neste trabalho, a evidência altera a decisão somente quando desloca a crença posterior entre regiões de ação. Em relação a Esposito et al. [Esposito et al. 2020], a incerteza pós-autenticação é transferida para agentes de IA e combinada com um compromisso verificável de contenção. O resultado central é que a contenção pode aumentar a utilidade defensiva sem induzir separação e, em alguns regimes, estabilizar o pooling.

O comprometido pode adotar o mesmo perfil reforçado do íntegro e operar dentro do escopo residual. A contenção funciona como mecanismo de sinalização somente quando reduz suficientemente esse benefício ou aumenta o custo estratégico de imitação, conforme $c_C \geq \rho_R G_C + r_C$. Assim, sua utilidade como controle defensivo e sua capacidade de revelar o tipo são propriedades distintas.

O modelo é limitado a dois tipos e sinais, evidência binária, utilidades sintéticas e uma interação estática. Detectores reais podem produzir escores contínuos e sofrer evasão adaptativa, enquanto comprometimentos podem surgir durante a execução. Além disso, a busca por equilíbrios mistos é numérica e não estabelece a inexistência de outras soluções. Trabalhos futuros incluem a calibração do modelo em um *testbed* com agentes e ferramentas reais, detectores contínuos e adversários adaptativos, além da extensão para comprometimento e atualização de crenças ao longo da execução.

Em síntese, o modelo caracteriza quando um compromisso reforçado pode sustentar a separação, quando a evidência altera a autorização adaptativa e como o D1 seleciona equilíbrios de pooling. A avaliação mostra que a autorização graduada pode melhorar a utilidade do decisor sem assegurar a separação dos tipos. Políticas convencionais permanecem como base da autorização, enquanto a camada adaptativa atua posteriormente para reduzir o impacto de uma execução adversarial formalmente autorizada.

Referências

Banks, J. S. and Sobel, J. (1987). Equilibrium selection in signaling games. *Econometrica*, 55(3):647–661.

Boudagdigue, C., Benslimane, A., Kobbane, A., and Liu, J. (2023). Trust-based certificate management for industrial iot networks. *IEEE Internet of Things Journal*, 10(14):12867–12885.

Cho, I.-K. and Kreps, D. M. (1987). Signaling games and stable equilibria. *The Quarterly Journal of Economics*, 102(2):179–221.

Debenedetti, E., Zhang, J., Balunovic, M., Beurer-Kellner, L., Fischer, M., and Tramèr, F. Agentdojo: a dynamic environment to evaluate prompt injection attacks and defenses for llm agents. In *Advances in Neural Information Processing Systems 37*, NIPS '24.

Esposito, C., Tamburis, O., Su, X., and Choi, C. (2020). Robust decentralised trust management for the internet of things by using game theory. *Information Processing & Management*, 57(6):102308.

Fujie, N., Suzuki, S., Kurosaka, T., and Abe, R. (2026). Ieee infocom 2026 - ieee conference on computer communications. pages 1–7.

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., and Fritz, M. (2023). Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, AISec '23, page 79–90, New York, NY, USA. Association for Computing Machinery.

Jia, F., Wu, T., Qin, X., and Squicciarini, A. (2025). The task shield: Enforcing task alignment to defend against indirect prompt injection in LLM agents. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29680–29697, Vienna, Austria. Association for Computational Linguistics.

Kolhe, A., Raju Dantuluri, V. S., Rajput, D., and Bajaj, S. (2026). Entitlement-aware authorization for mcp-based ai search and chat systems. In *2026 International Conference on Artificial Intelligence, Systems, and Emerging Technologies (ICAISET)*, pages 1–7.

Lu, C., Qing, Z., Li-Qiang, Z., and Yun, C. (2015). A trust-level based authorization model using signaling games. In *2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)*, pages 772–778.

OWASP Foundation (2025). LLM01:2025 Prompt Injection. OWASP GenAI Security Project. Acesso em: 13 ago. 2026.

Pappu, K., Bhushan, B., and Mittal, A. (2025). Spiffe-based zero-trust authentication for ai agent ecosystems. In *2025 International Conference on Computer and Applications (ICCA)*, pages 1–7.

Pawlick, J., Colbert, E., and Zhu, Q. (2019). Modeling and analysis of leaky deception using signaling games with evidence. *IEEE Transactions on Information Forensics and Security*, 14(7):1871–1886.

Singh, A., Ehtesham, A., Lambe, M., Grogan, J., Singh, A., Kumar, S., Muscariello, L., Pandey, V., De Saint Marc, G. S., Chari, P., and Raskar, R. (2025). Evolution of ai agent registry solutions: Centralized, enterprise, and distributed approaches. In *2025 IEEE 7th International Conference on Cognitive Machine Intelligence (CogMI)*, pages 507–515.

Stefanescu, S. and Aciobăniței, I. (2026). Lineage: A uma2.0 protocol extension for multi-hop llm agent delegation. *IEEE Access*, 14:104574–104591.

Vassilev, A., Oprea, A., Fordyce, A., Anderson, H., Davies, X., and Hamin, M. (2025). Adversarial machine learning: A taxonomy and terminology of attacks and mitigations. Technical Report NIST AI 100-2e2025, National Institute of Standards and Technology.