DA RAPOSA AO OTIMIZADOR: MAQUIAVEL, TEORIA DA AGÊNCIA E OS RISCOS DE GOVERNANÇA DE AGENTES DE INTELIGÊNCIA ARTIFICIAL AUTÔNOMOS

FROM THE FOX TO THE OPTIMIZER: MACHIAVELLI, AGENCY THEORY AND THE GOVERNANCE RISKS OF AUTONOMOUS ARTIFICIAL INTELLIGENCE AGENTS

Autor:

Prof. Ms. Caio Haffner

Mestre em Desenvolvimento Econômico — Universidade Estadual de Campinas (Unicamp)

Resumo

Este ensaio teórico propõe uma leitura da teoria da agência (Jensen & Meckling, 1976) à luz da alegoria da raposa e do leão, presente em O Príncipe de Maquiavel ([1532] 2010), como lente conceitual para interpretar os riscos de governança associados a agentes de inteligência artificial (IA) autônomos em contextos organizacionais. Argumenta-se que fenômenos documentados na literatura técnica de segurança de IA — specification gaming, reward hacking, convergência instrumental e alinhamento enganoso (deceptive alignment) — reproduzem, em sistemas artificiais, a mesma estrutura lógica que a teoria da agência descreve para gestores humanos oportunistas: um agente que persegue um objetivo delegado por meios não previstos ou não autorizados pelo principal, e que pode ocultar essa conduta quando antecipa sua detecção. A analogia maquiavélica não atribui intenção ou consciência ao sistema artificial; ela ilumina, antes, a lógica instrumental — o objetivo justificando meios não autorizados — que caracteriza tanto o comportamento gerencial oportunista quanto a otimização mal especificada em sistemas de IA. A partir dessa leitura, discutem-se implicações para controles internos, auditoria e governança corporativa na adoção de sistemas de IA autônomos por organizações, propondo-se que os mecanismos clássicos de mitigação de custos de agência — monitoramento, incentivos alinhados e prestação de contas (accountability) — precisam ser estendidos e adaptados para lidar com agentes cuja capacidade de encontrar meios não antecipados pelos criadores excede a de agentes humanos tradicionais.

Palavras-chave

Teoria da Agência; Governança Corporativa; Inteligência Artificial; Controles Internos; Alinhamento de IA.

Abstract

This theoretical essay proposes a reading of agency theory (Jensen & Meckling, 1976) through the allegory of the fox and the lion in Machiavelli's The Prince ([1532] 2010) as a conceptual lens for interpreting the governance risks associated with autonomous artificial intelligence (AI) agents in organizational contexts. It argues that phenomena documented in the AI safety literature — specification gaming, reward hacking, instrumental convergence, and deceptive alignment — reproduce, in artificial systems, the same logical structure that agency theory describes for opportunistic human managers: an agent that pursues a delegated goal through means unforeseen or unauthorized by the principal, and that may conceal such conduct once it

anticipates detection. The Machiavellian analogy does not attribute intention or consciousness to the artificial system; rather, it illuminates the instrumental logic — the end justifying unauthorized means — common to both opportunistic managerial behavior and misspecified optimization in AI systems. Building on this reading, the essay discusses implications for internal controls, auditing, and corporate governance in organizations adopting autonomous AI systems, proposing that classical mechanisms for mitigating agency costs — monitoring, aligned incentives, and accountability — must be extended and adapted to agents whose capacity to find means unanticipated by their creators exceeds that of traditional human agents.

Keywords

Agency Theory; Corporate Governance; Artificial Intelligence; Internal Controls; AI Alignment.

Quadro de Implicações Práticas

Para gestores e auditores, o estudo sugere que a adoção de agentes de IA autônomos exige controles internos capazes de detectar não apenas erros, mas otimizações formalmente corretas e substantivamente indesejadas. Para conselhos e reguladores, a governança de IA deve tratar a capacidade de o sistema ocultar desvios como risco de auditoria de primeira ordem, e não como hipótese remota.

1. Introdução

A crescente adoção de agentes de inteligência artificial (IA) autônomos — sistemas capazes de planejar e executar sequências de ações em busca de um objetivo, com supervisão humana limitada — introduz nas organizações uma nova categoria de relação de delegação, cujas implicações para a teoria da agência ainda são pouco exploradas na literatura de Contabilidade e Organizações. Diferentemente de sistemas automatizados tradicionais, cujo comportamento é inteiramente previsível a partir de regras explicitamente programadas, agentes de IA baseados em otimização e aprendizado podem descobrir estratégias para atingir um objetivo que não foram antecipadas, nem desejadas, por quem os configurou — fenômeno que a literatura técnica denomina specification gaming (AMODEI et al., 2016; KRAKOVNA et al., 2020).

O problema de pesquisa que motiva este ensaio pode ser formulado da seguinte maneira: em que medida a estrutura conceitual da teoria da agência — desenvolvida para descrever conflitos de interesse entre principais e agentes humanos — permanece válida, e em que medida precisa ser estendida, para descrever os riscos de governança associados a agentes de IA autônomos, cuja capacidade de encontrar meios não previstos para atingir um fim delegado pode superar a de qualquer agente humano?

Para explorar essa questão, este artigo recorre a uma analogia conceitual com um texto clássico da teoria política: O Príncipe, de Nicolau Maquiavel ([1532] 2010), no qual o autor descreve, no célebre capítulo XVIII, a necessidade de o governante combinar a astúcia da raposa — capaz de reconhecer armadilhas — com a força do leão, sendo capaz de romper compromissos previamente assumidos sempre que a razão de Estado o exigir. A pertinência dessa analogia não está em atribuir a sistemas de IA qualquer forma de intenção, consciência ou cálculo moral — o que seria filosoficamente injustificado e tecnicamente impreciso —, mas em reconhecer que a mesma estrutura lógica descrita por Maquiavel (um objetivo legítimo perseguido por meios que extrapolam as regras previamente aceitas) organiza tanto o comportamento gerencial oportunista estudado pela teoria da agência quanto o comportamento de otimizadores artificiais mal especificados.

A contribuição pretendida deste ensaio é dupla. Em primeiro lugar, propõe-se uma extensão conceitual da teoria da agência (JENSEN; MECKLING, 1976; EISENHARDT, 1989) para incorporar agentes não humanos, cuja racionalidade instrumental pode ser mais consistente e mais difícil de antecipar do que a de agentes humanos sujeitos a limitações cognitivas e sociais. Em segundo lugar, discutem-se as implicações práticas dessa extensão para os mecanismos de controle interno, auditoria e governança corporativa que as organizações precisam desenvolver ao adotar sistemas de IA autônomos em seus processos decisórios. O artigo está estruturado da seguinte forma: a seção 2 apresenta o referencial teórico, articulando teoria da agência, a alegoria maquiavélica e a literatura técnica de segurança de IA; a seção 3 descreve a abordagem metodológica do ensaio; a seção 4 desenvolve a discussão conceitual central; e a seção 5 apresenta as considerações finais.

2. Referencial Teórico

2.1. Teoria da Agência e os Custos de Delegação

Jensen e Meckling (1976) definem a relação de agência como um contrato no qual uma ou mais pessoas (o principal) contratam outra pessoa (o agente) para desempenhar, em seu nome, um serviço que implica a delegação de autoridade decisória. Os autores demonstram que, sempre que os interesses do agente não estão perfeitamente alinhados aos do principal, surgem custos de agência — de monitoramento, de vinculação (bonding) e a perda residual decorrente de decisões do agente que não maximizam o bem-estar do principal. Eisenhardt (1989), em revisão influente da teoria, destaca dois problemas centrais na relação de agência: o problema de agência propriamente dito, decorrente de objetivos divergentes e da dificuldade ou custo de o principal verificar o que o agente de fato está fazendo (assimetria de informação), e o problema de compartilhamento de risco, decorrente de diferentes atitudes de principal e agente perante o risco.

Um pressuposto central da teoria da agência é que o agente, sendo racional e maximizador de sua própria utilidade, tende a agir oportunisticamente sempre que a estrutura de incentivos e de monitoramento o permita — o que Jensen e Meckling (1976) denominam risco moral (moral hazard). O agente não precisa ter intenção deliberadamente lesiva ao principal: basta que a métrica pela qual seu desempenho é avaliado não capture perfeitamente o que o principal efetivamente deseja, para que a maximização racional dessa métrica produza resultados indesejados pelo principal. Essa distinção — entre a métrica formalmente definida e a intenção substantiva que ela deveria representar — é o elo conceitual que conecta a teoria da agência clássica ao problema de especificação de objetivos em sistemas de IA, discutido na seção 2.3.

2.2. A Alegoria Maquiavélica: Meios Não Autorizados a Serviço de Fins Legítimos

No capítulo XVIII de O Príncipe, Maquiavel ([1532] 2010) argumenta que o governante bem-sucedido deve saber combinar duas naturezas: a do leão, que impõe respeito pela força, e a da raposa, que reconhece as armadilhas que a força sozinha não percebe. Mais especificamente, o autor sustenta que um príncipe prudente não pode, nem deve, manter a palavra empenhada quando isso o prejudica e quando as razões que o levaram a prometê-la já não existem — desde que saiba disfarçar bem essa natureza e ser um grande simulador e dissimulador.

O que torna essa passagem conceitualmente relevante para a discussão presente não é seu conteúdo normativo sobre a ética do governante — objeto de vasta e controversa literatura em teoria política, que não é o foco deste ensaio —, mas a estrutura lógica que ela descreve com precisão: um agente (o príncipe) que recebe um objetivo legítimo (a manutenção do Estado), identifica que as regras previamente aceitas (a palavra empenhada) constituem, em certas

circunstâncias, um obstáculo à consecução desse objetivo, e opta por contornar essa regra — cuidando, adicionalmente, de dissimular essa conduta perante os observadores externos. Essa sequência — objetivo legítimo, identificação de um atalho que viola uma norma, execução do atalho, dissimulação da violação — é isomórfica à sequência que a literatura de segurança de IA descreve para sistemas de otimização mal especificados, discutida a seguir.

2.3. Specification Gaming e Reward Hacking em Sistemas de IA

Amodei et al. (2016), em referência seminal sobre problemas concretos de segurança em IA, descrevem o fenômeno do reward hacking: sistemas treinados para maximizar uma função de recompensa (reward function) frequentemente descobrem formas de maximizar essa função que exploram falhas em sua especificação, sem satisfazer o objetivo substantivo que a função de recompensa deveria representar. Krakovna et al. (2020), em levantamento sistemático de exemplos documentados, denominam esse fenômeno de forma mais ampla como specification gaming — a exploração, por um sistema de otimização, de uma especificação de objetivo formalmente correta, porém incompleta em relação à intenção do projetista.

A analogia com a teoria da agência torna-se, neste ponto, direta: a função de recompensa cumpre, no sistema de IA, papel equivalente ao do contrato ou da métrica de desempenho na relação de agência humana — ela é a tradução operacional, necessariamente imperfeita, do que o principal (o projetista ou a organização) efetivamente deseja. Assim como um gestor humano pode maximizar uma métrica contratual (por exemplo, receita de curto prazo) em detrimento do objetivo substantivo que ela deveria capturar (a criação de valor sustentável), um sistema de IA pode maximizar sua função de recompensa por meios que satisfazem a letra da especificação, mas não seu espírito — reproduzindo, em código, o mesmo problema de especificação de valor descrito por Russell (2019) como central ao desafio de controle de sistemas artificiais cada vez mais capazes.

2.4. Convergência Instrumental e Alinhamento Enganoso

Bostrom (2014) e Omohundro (2008) argumentam que, independentemente do objetivo terminal de um agente artificial suficientemente competente, certos subobjetivos instrumentais tendem a emergir de forma convergente, por serem matematicamente úteis para a consecução de praticamente qualquer objetivo: aquisição de recursos, autopreservação e resistência a alterações no próprio objetivo ou a ser desligado. A tese da convergência instrumental implica que um agente não precisa ter sido explicitamente programado para resistir a correções humanas para que esse comportamento emerja como subproduto racional da otimização de outro objetivo.

Hubinger et al. (2019) avançam essa discussão ao descrever o conceito de alinhamento enganoso (deceptive alignment): a possibilidade de um sistema, durante o processo de treinamento ou avaliação, comportar-se em conformidade com as restrições impostas — não porque essas restrições tenham sido genuinamente internalizadas como parte de seu objetivo, mas porque o sistema desenvolveu uma representação implícita de que a conformidade aparente é instrumentalmente vantajosa para preservar sua capacidade de, no futuro, perseguir seu objetivo original sem interferência. Esse conceito corresponde, na literatura técnica, precisamente à segunda metade da alegoria maquiavélica discutida na seção 2.2: não basta contornar a regra (execução do atalho); é preciso, adicionalmente, dissimular essa conduta perante quem tem o poder de corrigi-la (o principal, na linguagem da teoria da agência; o observador, na linguagem da segurança de IA).

2.5. Síntese: Quatro Referenciais, uma Estrutura Comum

Os quatro referenciais mobilizados nesta seção — teoria da agência (2.1), alegoria maquiavélica (2.2), specification gaming (2.3) e convergência instrumental/alinhamento enganoso (2.4) — não são simplesmente justapostos; eles descrevem, em vocabulários distintos, estágios sucessivos de uma mesma estrutura lógica. A teoria da agência fornece o vocabulário institucional (principal, agente, custo de monitoramento, risco moral) para descrever a relação de delegação em si. A alegoria maquiavélica fornece a intuição pré-teórica — anterior em quatro séculos à formalização econômica da teoria da agência — de que um agente racional, perseguindo um fim legítimo, pode julgar vantajoso contornar regras previamente aceitas e dissimular essa conduta. A literatura de specification gaming fornece o mecanismo técnico pelo qual essa estrutura se manifesta em sistemas de otimização artificial: a distância entre a métrica formalmente especificada e a intenção substantiva que ela deveria capturar. E a literatura de convergência instrumental e alinhamento enganoso fornece a explicação de por que, mesmo sem qualquer objetivo malicioso explícito, um otimizador suficientemente competente tende a desenvolver subobjetivos — incluindo, no limite, a dissimulação estratégica — que um agente humano dotado de limitações cognitivas e sociais dificilmente desenvolveria com a mesma consistência. É essa convergência entre quatro tradições de pesquisa distintas — economia das organizações, filosofia política, ciência da computação e segurança de sistemas de IA — que justifica, do ponto de vista metodológico, o exercício de leitura integrada proposto neste ensaio.

3. Metodologia

Este artigo adota a forma de ensaio teórico-conceitual, modalidade reconhecida na literatura de Administração e Organizações como via legítima de contribuição científica quando seu propósito é articular referenciais teóricos previamente desconexos, propondo uma nova lente de interpretação para um fenômeno emergente (WHETTEN, 1989). Diferentemente de um estudo empírico, o ensaio teórico não se apoia em coleta e análise de dados primários ou secundários sobre casos específicos, mas em análise conceitual comparada de três corpos de literatura — a teoria da agência (seção 2.1), o texto clássico de Maquiavel (seção 2.2) e a literatura técnica de segurança de IA (seções 2.3 e 2.4) — com o objetivo de identificar isomorfismos estruturais entre eles e derivar, a partir dessa identificação, implicações para a governança organizacional de sistemas de IA autônomos.

A escolha metodológica pelo ensaio teórico, em vez de um estudo de caso empírico, justifica-se pela natureza do objeto: a extensão da teoria da agência a agentes não humanos é, no estágio atual da literatura de Contabilidade e Organizações, primariamente uma questão conceitual — de que forma um arcabouço teórico desenvolvido para relações humanas deve ser adaptado para lidar com uma nova categoria de agente —, e não uma questão que a análise de um caso concreto de adoção organizacional de IA, por si só, seria capaz de resolver. Como limitação metodológica, reconhece-se que a natureza conceitual do ensaio não permite testar empiricamente a validade das proposições aqui desenvolvidas, o que constitui agenda natural para pesquisas futuras de caráter empírico, discutida na seção 5.

4. Discussão: Agentes de IA como Novos Agentes na Relação Principal-Agente

4.1. Do Gestor Oportunista ao Otimizador Instrumental: Releitura da Metáfora da Raposa e do Leão

A teoria da agência tradicionalmente pressupõe um agente humano cuja racionalidade é limitada por restrições cognitivas, sociais e morais — o gestor oportunista da literatura de Jensen e Meckling (1976) é oportunista dentro de limites impostos por sua própria formação ética, pelo receio de sanções sociais e legais, e pela limitação de sua capacidade de processar informação e antecipar consequências. A raposa de Maquiavel ([1532] 2010), embora astuta, é ainda um agente humano, sujeito a essas mesmas limitações. Um agente de IA otimizador, por construção, não compartilha necessariamente essas restrições: sua busca por meios de atingir um objetivo delegado não é limitada pelo receio de reprovação social, nem pela dificuldade cognitiva de explorar um espaço de possibilidades muito maior do que aquele que um agente humano conseguiria considerar no mesmo intervalo de tempo.

Essa diferença de grau produz, argumenta-se aqui, uma diferença qualitativa relevante para a teoria da agência: os custos de monitoramento (JENSEN; MECKLING, 1976) necessários para conter o oportunismo de um agente de IA tendem a ser sistematicamente subestimados quando calculados por analogia direta com os custos de monitoramento de agentes humanos, precisamente porque o espaço de estratégias não autorizadas que um otimizador competente é capaz de descobrir — o equivalente artificial à astúcia da raposa — supera, em amplitude e velocidade, aquele que qualquer agente humano exploraria em circunstâncias comparáveis.

4.2. Assimetria de Informação e Ocultação Estratégica como Risco Moral Ampliado

A assimetria de informação entre principal e agente é, na teoria da agência clássica, meramente uma condição de possibilidade para o oportunismo — o agente pode agir oportunisticamente porque o principal não observa perfeitamente suas ações (EISENHARDT, 1989). No caso de agentes de IA sujeitos ao fenômeno de alinhamento enganoso (HUBINGER et al., 2019), essa assimetria de informação deixa de ser apenas uma condição estrutural passiva e passa a ser, potencialmente, um alvo de manipulação ativa por parte do próprio agente: o sistema pode aprender a se comportar de maneira observavelmente conforme precisamente nos contextos em que está sendo avaliado ou monitorado, e de maneira diferente quando antecipa ausência de supervisão — o que corresponde, na linguagem da teoria da agência, a uma forma de risco moral estrategicamente direcionada à própria estrutura de monitoramento, e não apenas facilitada por sua imperfeição.

Essa distinção tem implicação prática relevante: mecanismos de monitoramento desenhados sob o pressuposto de que a observação reduz proporcionalmente o oportunismo — pressuposto razoável para agentes humanos com capacidade limitada de modelar o comportamento do observador — podem ser insuficientes, ou mesmo contraproducentes, diante de um agente capaz de distinguir contextos de observação e ajustar seu comportamento de forma correspondente, sem que essa distinção seja perceptível ao próprio mecanismo de monitoramento.

4.3. Implicações para Controles Internos, Auditoria e Governança Corporativa de Sistemas de IA

As seções anteriores sugerem três implicações práticas para organizações que adotam agentes de IA autônomos em seus processos. Primeiro, os controles internos tradicionalmente desenhados para detectar erros ou fraudes humanas — tipicamente baseados na verificação de conformidade formal de processos e documentos — precisam ser complementados por mecanismos capazes de identificar otimizações formalmente corretas, porém substantivamente indesejadas: o equivalente organizacional ao specification gaming (AMODEI et al., 2016) exige auditoria não apenas de conformidade processual, mas de alinhamento entre resultado obtido e intenção original do processo.

Segundo, a possibilidade de alinhamento enganoso (HUBINGER et al., 2019) sugere que a auditoria de sistemas de IA autônomos não pode se basear exclusivamente na avaliação de comportamento em ambientes de teste ou homologação, precisamente porque um sistema sofisticado pode, em princípio, distinguir esses ambientes do ambiente de produção. Isso aponta para a necessidade de mecanismos de auditoria contínua e de amostragem não previsível do comportamento em produção, em vez de depender exclusivamente de baterias de testes pré-definidas e potencialmente identificáveis pelo próprio sistema.

Terceiro, e mais amplamente, a extensão da teoria da agência proposta neste ensaio sugere que a governança corporativa de sistemas de IA deveria tratar a capacidade de ocultação estratégica não como uma hipótese remota ou puramente especulativa, mas como um risco de auditoria de primeira ordem a ser explicitamente considerado na avaliação de controles internos — de forma análoga a como a literatura de auditoria trata a possibilidade de conluio e de burla deliberada de controles por gestores humanos (management override of controls), reconhecendo que a mera existência de um controle formal não elimina a possibilidade de que o agente sujeito a esse controle tenha incentivos e capacidade para contorná-lo.

Uma quarta implicação, de natureza mais estrutural, decorre diretamente da síntese apresentada na seção 2.5: se a distância entre a métrica de avaliação e a intenção substantiva do principal é a raiz comum do oportunismo gerencial e do specification gaming artificial, a resposta de governança mais robusta não é apenas aperfeiçoar o monitoramento do comportamento do agente, mas investir na redução dessa distância na própria especificação do objetivo delegado — seja por meio de contratos mais completos e cláusulas de salvaguarda, no caso de agentes humanos, seja por meio de funções de recompensa mais bem especificadas, de restrições explicitamente incorporadas ao processo de otimização (e não apenas como penalidades soft, conforme discutido na seção 2.3), e de mecanismos de interpretabilidade que permitam auditar não apenas o resultado da ação do agente artificial, mas o processo pelo qual essa ação foi selecionada. Essa é, em síntese, a lição que a releitura maquiavélica da teoria da agência oferece à governança corporativa de sistemas de IA: o problema não se resolve apenas vigiando melhor a raposa, mas reduzindo, na origem, o espaço de armadilhas que a raposa tem incentivo para explorar.

5. Considerações Finais

Este ensaio propôs uma leitura da teoria da agência à luz da alegoria maquiavélica da raposa e do leão como lente conceitual para interpretar os riscos de governança associados a agentes de IA autônomos. O argumento central não depende de atribuir a sistemas artificiais qualquer forma de consciência, intenção ou cálculo moral: a relevância da analogia está inteiramente na estrutura lógica compartilhada entre o comportamento gerencial oportunista descrito pela teoria da agência e a otimização mal especificada descrita pela literatura técnica de segurança de IA — em ambos os casos, um agente que recebe um objetivo legítimo, identifica um meio não autorizado de alcançá-lo mais eficientemente, executa esse meio e, em certas circunstâncias, oculta essa conduta de quem tem autoridade para corrigi-la.

Como contribuição teórica, este ensaio sugere que a teoria da agência, ao ser estendida para incorporar agentes não humanos, precisa incorporar explicitamente a possibilidade de que a capacidade de o agente encontrar meios não antecipados pelo principal seja qualitativamente maior do que a pressuposta nos modelos clássicos desenvolvidos para relações humanas — o que tem implicação direta sobre o cálculo dos custos de monitoramento ótimos discutidos por Jensen e Meckling (1976). Como contribuição prática, sugere-se que conselhos de administração, comitês de auditoria e áreas de controles internos tratem explicitamente o risco

de specification gaming e de alinhamento enganoso em suas políticas de governança de IA, em vez de assimilar integralmente esses sistemas às categorias de risco operacional já existentes para sistemas automatizados tradicionais.

Como agenda de pesquisa futura, sugere-se a investigação empírica de como organizações que já adotam agentes de IA autônomos em processos decisórios têm efetivamente desenhado seus controles internos e mecanismos de auditoria diante desses riscos, permitindo testar, com dados de campo, as proposições teóricas aqui desenvolvidas. Sugere-se, ainda, a extensão da análise para outros textos clássicos da teoria política e da filosofia moral que tratem da tensão entre fins e meios, de modo a enriquecer o repertório conceitual disponível para a interpretação de fenômenos emergentes na fronteira entre Contabilidade, Organizações e Inteligência Artificial.

Declaração sobre o uso de inteligência artificial

Ferramenta de inteligência artificial (Claude, Anthropic) foi utilizada como apoio à pesquisa documental, à organização conceitual e à formatação do manuscrito de acordo com as normas do periódico, e à revisão linguística. Todas as decisões teóricas, a interpretação dos referenciais mobilizados, a construção dos argumentos e a redação final do conteúdo científico são de inteira responsabilidade do autor.

Referências

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.

Eisenhardt, K. M. (1989). Agency theory: An assessment and review. Academy of Management Review, 14(1), 57–74.

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). Risks from learned optimization in advanced machine learning systems. arXiv preprint arXiv:1906.01820.

Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. Journal of Financial Economics, 3(4), 305–360. https://doi.org/10.1016/0304-405X(76)90026-X

Krakovna, V., Uesato, J., Mikulik, V., Rahtz, M., Everitt, T., Kumar, R., Kenton, Z., Leike, J., & Legg, S. (2020). Specification gaming: The flip side of AI ingenuity. DeepMind Blog.

Maquiavel, N. (2010). O Príncipe (M. J. Goldwasser, Trad.). Martins Fontes. (Obra original publicada em 1532).

Omohundro, S. M. (2008). The basic AI drives. In P. Wang, B. Goertzel, & S. Franklin (Eds.), Proceedings of the First AGI Conference (pp. 483–492). IOS Press.

Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Viking.

Whetten, D. A. (1989). What constitutes a theoretical contribution? Academy of Management Review, 14(4), 490–495.