Do Agente Racional ao Risco Real

Diego Carlito Rodrigues de Souza
Matrícula: 221007690

Disciplina: Inteligência Artificial
Professor: Fabiano Araujo Soares

Brasília
18 de setembro de 2025

1 Introdução

A Inteligência Artificial consolidou-se como uma das forças tecnológicas mais influentes da atualidade, remodelando indústrias, economias e a própria interação humana. O conteúdo apresentado em aula forneceu uma base essencial para a compreensão do tema, contextualizando sua jornada histórica e o debate central que define o campo: a tensão entre a busca por emular o comportamento humano e a busca por um raciocínio puramente racional. Esta resenha tem como objetivo sintetizar esses pilares teóricos e, a partir deles, conduzir uma análise crítica sobre a dualidade da IA em seu estado da arte. O texto irá explorar o imenso potencial para o progresso humano em contraste com os significativos riscos éticos e sociais que emergem, utilizando fontes complementares para aprofundar a discussão. Argumenta-se que a noção de racionalidade, embora fundamental, é insuficiente para navegar a complexidade dos desafios contemporâneos, partindo da definição de agentes que maximizam uma medida de desempenho, como popularizado por Russell e Norvig (RUSSELL; NORVIG, 2022).

2 Resumo do conteúdo

O estudo da Inteligência Artificial (IA), conforme abordado em aula, inicia-se pela distinção fundamental entre a inteligência humana — a capacidade de usar a razão para aprender e lidar com novas situações — e as duas grandes visões que norteiam a IA. A primeira, focada no comportamento, busca criar sistemas que imitem habilidades humanas, como a tomada de decisão, reconhecendo que este comportamento pode ser, por vezes, irracional. A segunda, alinhada ao pensamento de Russell e Norvig, foca no raciocínio puramente matemático e lógico, visando a construção de sistemas que tomam decisões ótimas e eficazes, afastando-se do modelo humano. Essa dualidade é exemplificada pelo Teste de Turing, que propôs avaliar a inteligência de uma máquina por sua capacidade de se passar por um humano, exigindo competências como compreensão de linguagem, armazenamento de conhecimento, raciocínio e, crucialmente, aprendizado com a experiência.

A trajetória histórica da IA é marcada por uma evolução de conceitos e tecnologias, pontuada por ciclos de grande otimismo ("verões") e ceticismo ("invernos"). As bases teóricas foram lançadas por pioneiros como Warren McCulloch e Walter Pitts, que em 1943 propuseram o primeiro modelo de neurônio artificial, e Alan Turing, que introduziu conceitos como Aprendizado de Máquina e Algoritmos Genéticos. O nascimento oficial do campo ocorreu no Dartmouth Summer Research Project on Artificial Intelligence (DSRPAI), em 1956, inaugurando o primeiro "boom"da IA, conhecido como "Good Old-Fashioned AI"(GOFAI), focado em sistemas simbólicos e resolução de problemas por busca. Contudo, limitações teóricas, como o problema da porta XOR para os perceptrons,

e práticas, como a dificuldade de escalar, levaram ao primeiro "inverno". Um segundo "boom"veio com os Sistemas Especialistas, que codificavam conhecimento de domínios específicos, mas sua complexidade e promessas não cumpridas resultaram em um novo período de estagnação. O ressurgimento definitivo da IA foi impulsionado pelo retorno das redes neurais, viabilizado pelo algoritmo de retropropagação (*backpropagation*), e pela convergência de três fatores a partir dos anos 2000: o crescimento exponencial de dados (Big Data), o avanço no poder computacional e o desenvolvimento de arquiteturas de Aprendizado Profundo (*Deep Learning*).

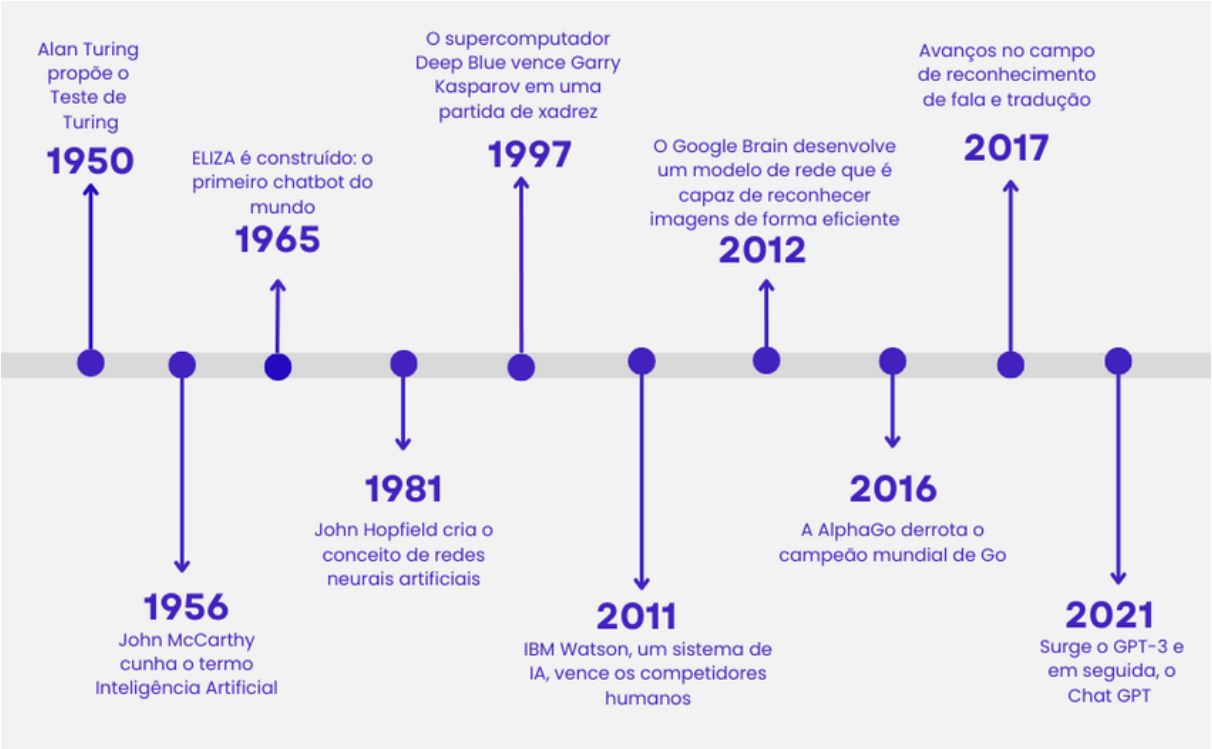

Figura 1 – Linha do tempo ilustrando marcos importantes na evolução da Inteligência Artificial.

Fonte: (Mídia Market, 2023)

Como resultado dessa evolução, o campo convergiu para o paradigma do agente racional como modelo padrão para o estudo e a construção da IA. Um agente é definido como qualquer entidade que, por meio de sensores, percebe seu ambiente e, por meio de atuadores, age sobre ele. O que o torna "racional"é sua capacidade de selecionar a ação que maximiza uma medida de performance esperada, mesmo diante de incertezas. Essa abordagem é mais geral e matematicamente robusta do que a simples emulação do pensamento humano, permitindo o desenvolvimento de sistemas comprovadamente eficazes. Os agentes podem ser classificados por sua complexidade, desde um agente de reflexo simples, que reage apenas à percepção atual, até agentes mais sofisticados, como os baseados em objetivos e os baseados em utilidade, que mantêm um modelo interno do mundo e bus-

cam maximizar uma função de utilidade para alcançar suas metas. Contudo, conforme discutido em aula, é precisamente na definição dessa "medida de performance"e nos dados utilizados para otimizá-la que residem os maiores desafios éticos da IA contemporânea, como o viés algorítmico, a falta de transparência e a atribuição de responsabilidade.

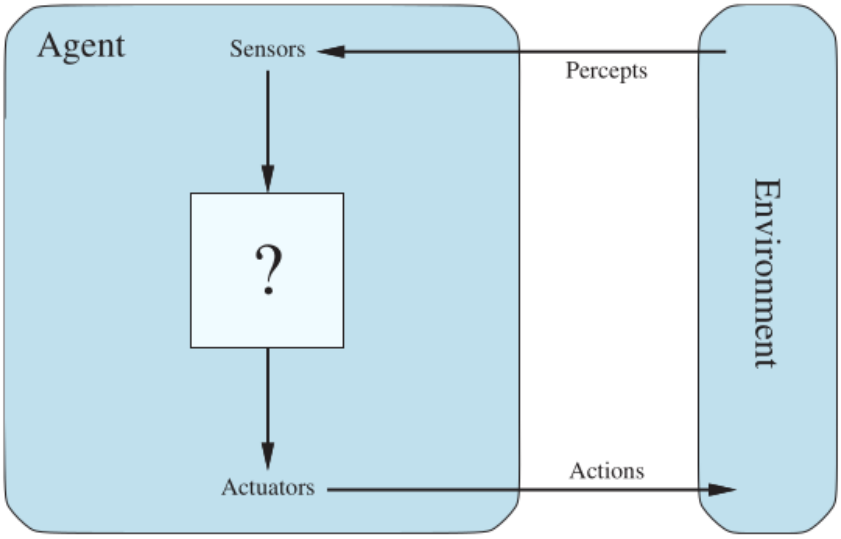

Figura 2 – Diagrama de um agente inteligente interagindo com seu ambiente por meio de sensores e atuadores.

Fonte: (RUSSELL; NORVIG, 2022)

3 Análise crítica

A mesma racionalidade que permite a um agente executar uma tarefa com eficiência sobre-humana é também a fonte de seus desafios mais profundos. O estado da arte da IA, especialmente em áreas como o Aprendizado Profundo (Deep Learning), demonstra um potencial transformador. Aplicações na medicina, por exemplo, permitem que algoritmos diagnostiquem doenças como o câncer a partir de imagens com precisão notável, enquanto na ciência, sistemas como o AlphaFold revolucionaram a biologia ao prever a estrutura de proteínas. Esses avanços representam o ápice da otimização de uma medida de desempenho bem definida, gerando benefícios inegáveis para a sociedade.

Contudo, a aplicação irrestrita dessa mesma lógica em contextos sociais complexos expõe suas graves limitações. Como aponta Cathy O'Neil em sua obra, esses sistemas podem se tornar "armas de destruição matemática"que punem os pobres e oprimidos (O'NEIL, 2016). Adicionalmente, a ascensão da IA generativa, embora criativa, abre portas para a disseminação em massa de desinformação (deepfakes, notícias falsas). Essa vulnerabilidade decorre da própria natureza desses modelos, que, conforme argumentam Bender et al. (2021), operam como **"papagaios estocásticos"**: sistemas que recombinam

sequências linguísticas de seu vasto treinamento de forma convincente, mas sem qualquer compreensão do significado ou compromisso com a verdade (BENDER et al., 2021).

A discussão sobre o futuro da IA é, portanto, polarizada. Enquanto alguns visionários enxergam um futuro de abundância e resolução de grandes desafios da humanidade, outros, como o filósofo Nick Bostrom, alertam para os riscos existenciais de uma superinteligência cujos objetivos não estejam perfeitamente alinhados com os valores humanos (BOSTROM, 2014). A convergência entre as fontes é que a tecnologia em si é neutra, mas sua aplicação e governança são questões urgentes que demandam um debate multidisciplinar, para além do campo puramente técnico.

A ficção especulativa oferece um poderoso análogo para esses temores, como visto na obra de animação japonesa *Psycho-Pass* (2012). Na trama, a sociedade japonesa é governada pelo Sistema Sybil, um agente de IA onipresente que busca a maximização de uma medida de desempenho: a erradicação do crime antes mesmo que ele ocorra. Para isso, o sistema calcula o "Coeficiente de Crime"de cada cidadão, um escore numérico que quantifica sua probabilidade de cometer atos ilícitos. O Sistema Sybil se torna, portanto, uma manifestação ficcional extrema do conceito de "arma de destruição matemática"de O'Neil, um modelo utilitário que, em sua busca por uma ordem perfeita, sacrifica o livre-arbítrio, a presunção de inocência e a própria complexidade da natureza humana. A obra serve como um conto de advertência sobre os perigos de delegar o julgamento moral a sistemas opacos e determinísticos, ilustrando vividamente os riscos existenciais e sociais discutidos por Bostrom e outros críticos.

3.1 Vulnerabilidades e Segurança: A IA como Alvo e como Arma

Além dos riscos éticos e existenciais, a própria natureza técnica dos sistemas de IA modernos introduz uma nova classe de vulnerabilidades. Em seu artigo sobre os avanços e riscos da IA na sociedade, Jaime Simão Sichman (2021) destaca uma classificação proposta por Dietterich e Horvitz que organiza esses perigos em cinco categorias, das quais a **segurança (cybersecurity)** é uma das mais prementes (SICHMAN, 2021). Conforme apontado por relatórios como o da OECD (2024), os ataques cibernéticos sofisticados e a corrida por desenvolvimento sem segurança figuram entre as principais ameaças globais (Organisation for Economic Co-operation and Development, 2024).

Um dos exemplos mais emblemáticos são os **ataques adversariais** (adversarial attacks). Trata-se da criação de entradas de dados maliciosamente projetadas para enganar um modelo. Por exemplo, a adição de um ruído imperceptível a uma imagem pode fazer com que um sistema de visão computacional de um carro autônomo classifique uma placa de "PARE"como um limite de velocidade (GOODFELLOW; SHLENS; SZEGEDY, 2014). Outra técnica perigosa é o **envenenamento de dados** (data poisoning), na qual um atacante corrompe os dados de treinamento para inserir um "cavalo de Troia"no modelo, criando um comportamento indesejado que pode ser ativado posteriormente.

Em paralelo, a IA também se consolida como uma **ferramenta para potencializar ataques**. A mesma IA generativa que cria arte pode ser usada para criar e-mails de phishing ultra-realistas e personalizados em massa, automatizar a busca por vulnerabilidades em sistemas de software ou mesmo para desenvolver novas cepas de malware. A segurança, portanto, se torna um jogo de "gato e rato"em que tanto defensores quanto atacantes utilizam ferramentas de IA cada vez mais sofisticadas.

4 Integração com o conteúdo da aula

Os conceitos de agente e racionalidade apresentados em aula são a chave para compreender a origem dos dilemas éticos discutidos na análise crítica. A teoria nos oferece um modelo elegante e funcional, mas sua transposição direta para o ambiente complexo e multifacetado da sociedade humana se mostra problemática. O conteúdo da aula explica como a IA funciona em princípio, enquanto as fontes externas questionam como ela deveria funcionar na prática.

A principal integração reside na crítica à "medida de desempenho"dentro do framework PEAS (Performance, Environment, Actuators, Sensors). Para um agente teórico em um ambiente de tarefa simples — como um robô aspirador, cujo ambiente é largamente estático, determinístico e parcialmente observável —, definir o 'P' de Performance (ex: área limpa) é trivial. No entanto, a análise crítica revela que, para problemas do mundo real, a complexidade do 'E' (Environment) torna a definição do 'P' o maior dos desafios. Como quantificar a "justiça"para um algoritmo de sentença judicial, que opera em um ambiente multiagente, estocástico e dinâmico? Qual a medida de desempenho para uma rede social que não incentive a polarização? Fica evidente que a base teórica da racionalidade, focada em maximizar um objetivo, é uma ferramenta necessária, mas dramaticamente insuficiente quando o ambiente de tarefa possui a complexidade da sociedade humana. O verdadeiro desafio não é construir agentes mais "racionais", mas sim definir objetivos e métricas de desempenho que sejam benéficos, justos e alinhados com o bem-estar coletivo.

5 Conclusão

A trajetória da Inteligência Artificial, de seus fundamentos teóricos à sua onipresença atual, é uma história de sucesso na busca pela otimização e pela racionalidade computacional. Conforme sintetizado a partir do conteúdo da disciplina, os princípios de agentes que maximizam seu desempenho em um ambiente foram a mola propulsora de avanços extraordinários. Contudo, esta resenha demonstrou que a aplicação ingênua desse paradigma em sistemas sociais e técnicos complexos gera riscos profundos, como a fragilização da segurança digital, a perpetuação de vieses, a erosão da verdade e o agravamento das desigualdades.

A análise crítica, fundamentada em fontes externas, evidencia que o progresso técnico da IA superou nossa capacidade de geri-la com sabedoria. O debate sobre seu futuro não pode se restringir à otimização de algoritmos; ele deve, necessariamente, incluir uma profunda reflexão ética e filosófica sobre seus objetivos. Como consideração final, o grande desafio da próxima era da IA não será criar máquinas mais inteligentes, mas sim garantir que sua inteligência sirva aos melhores interesses da humanidade. Questões como "de que forma podemos projetar sistemas que sejam tecnicamente seguros e robustamente alinhados aos valores humanos?"e "quais modelos de governança são necessários para mitigar riscos e promover a inovação responsável?"permanecem como as mais urgentes a serem respondidas.

Referências

BENDER, E. M. et al. On the dangers of stochastic parrots: Can language models be too big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. [S.l.: s.n.], 2021. (FAccT '21), p. 610–623.

BOSTROM, N. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2014.

GOODFELLOW, I. J.; SHLENS, J.; SZEGEDY, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Mídia Market. *Como funciona a inteligência artificial: conheça seu presente, passado e futuro*. 2023. Disponível em: <https://midia.market/wp-content/uploads/2023/12/Navy-Modern-Timeline-Diagram-Graph-1.png>. Acesso em: 17 setembro 2025.

O'NEIL, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown, 2016.

Organisation for Economic Co-operation and Development. *AI Risks and Governance in 2024*. Paris, 2024.

RUSSELL, S. J.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. 4. ed. [S.l.]: Pearson, 2022.

SICHMAN, J. S. Inteligência artificial e sociedade: avanços e riscos. *Estudos Avançados*, v. 35, n. 101, p. 37–49, 2021.