PONTIFÍCIA UNIVERSIDADE CATÓLICA DE SÃO PAULO

A ORIGEM E OS DESTINOS DA INTENCIONALIDADE:

Investigação da intencionalidade na pré-história e dos desenvolvimentos da intencionalidade artificial

SÃO PAULO

2025

André Magno

Trabalho de Qualificação a ser apresentado à Banca Examinadora da Pontifícia Universidade Católica de São Paulo, como exigência parcial para obtenção do título de MESTRE em Tecnologias da Inteligência e Design Digital – TIDD, sob orientação da Profa. Dra. Dora Kaufman e coorientação do Prof. Dr. João Carlos Moreno.

SÃO PAULO
2025

Banca Examinadora

_____

Profa. Dra. Dora Kaufman (presidente) PUC-SP

_____

Profa. Dra. Lucia Santaella

_____

Profa. Dra. Leticia Cristina Correa

RESUMO

A intencionalidade – capacidade da mente de se direcionar a objetos, ideias e ações – constitui um traço essencial da cognição humana, articulando razão, cultura e materialidade. Esta dissertação empreende uma investigação que atravessa a pré-história, examinando vestígios arqueológicos de intencionalidade consciente até alcançar os avanços contemporâneos da Inteligência Artificial, que tensionam as fronteiras entre simulação e agência real. A partir de um diálogo entre filosofia da mente, arqueologia cognitiva e ciência da computação, busca-se compreender se a intencionalidade pode emergir em sistemas algorítmicos ou se permanece um atributo irredutível da consciência biológica. A análise percorre autores reconhecidos, integrando evidências empíricas e teorias sobre cognição, técnica e agência. Ao investigar a gênese e os possíveis destinos da intencionalidade, a pesquisa oferece uma reflexão crítica sobre os limites da mente humana e os horizontes de sua replicação técnica.

Palavras-chave: Intencionalidade; Arqueologia Cognitiva; Inteligência Artificial; Intencionalidade artificial; Cognição e Tecnologia; Agência Técnica

ABSTRACT

Intentionality – the mind's capacity to be directed toward objects, ideas, and actions – is a defining trait of human cognition, interweaving reason, culture, and materiality. This dissertation explores a trajectory from the earliest archaeological traces of conscious intentionality to the contemporary developments in artificial intelligence, which challenge the boundaries between simulation and genuine agency. Through an interdisciplinary dialogue between philosophy of mind, cognitive archaeology, and computer science, the study seeks to understand whether intentionality can emerge in algorithmic systems or remains irreducibly tied to biological consciousness. The analysis engages key thinkers, integrating empirical evidence and theoretical insights on cognition, technology, and agency. By investigating the origins and potential futures of intentionality, this research offers a critical reflection on the limits of the human mind and the horizons of its technical replication.

Keywords: Intentionality; Cognitive Archaeology; Artificial Intelligence; Algorithmic Intentionality; Cognition and Technology; Technological Agency

SUMÁRIO

INTRODUÇÃO 8

1.1 Motivação da pesquisa: uma trajetória profissional na fronteira da inovação 8

1.2 O racional da pesquisa: uma jornada em três atos 9

1.3 Metodologia: fundamentos de uma investigação interdisciplinar 11

1.4 O que é Inteligência Artificial? Desafios de uma definição 13

1.5 A genealogia da Inteligência Artificial: uma trajetória histórica 15

1.6 A mecânica dos modelos multifuncionais: a arquitetura *transformer* 20

1.7 Rumo à autonomia: a emergência das IAs Agênticas 22

2. A INTENCIONALIDADE 23

2.1. O Naturalismo Biológico: a intencionalidade como fenômeno causal do cérebro 24

2.2. Intencionalidade intrínseca vs. derivada e "como-se" 25

2.3. A estrutura da intencionalidade intrínseca 26

2.4. Estruturas de suporte: a rede e o *background* 28

2.5. A conexão indissociável: intencionalidade, consciência e cognição 29

2.6 A perspectiva evolutiva: a origem biológica e os níveis da intencionalidade 32

2.7. O desafio da Inteligência Artificial forte e o Argumento do Quarto Chinês 34

2.8 O contraponto funcionalista: a "atitude intencional" de Daniel Dennett 35

2.9 Por que prevalecer com a teoria de Searle 36

2.10. Conclusão do capítulo 38

3. ORIGENS DA INTENCIONALIDADE: UMA INVESTIGAÇÃO ARQUEOLÓGICA 39

3.1. A arqueologia da mente: inferindo a intenção a partir da pedra 40

3.2. As primeiras evidências: da percussão oportunista à ferramenta sistematizada 41

3.3. A consolidação da intencionalidade: simetria e padronização no Acheulense 47

3.4. A dimensão social da mente: intencionalidade individual vs. compartilhada 51

3.5. Da função à origem do significado: uma perspectiva paleoantropológica 56

3.6. A mente no cérebro: evidências da neuroarqueologia e a coevolução da tecnologia e da linguagem.....57

3.7. O motor da mente: a hipótese do tecido caro e a dieta.....62

4. A EMERGÊNCIA DA INTENCIONALIDADE EM SISTEMAS ARTIFICIAIS – HORIZONTES E LIMITES 65

4.1 A genealogia da intencionalidade derivada: uma história crítica da Inteligência Artificial 67

4.2. A barreira da causalidade: por que a simulação não se torna intencionalidade ...69

4.3. O debate central sobre a intencionalidade intrínseca em substratos não biológicos 71

4.4 Panorama das pesquisas atuais: um mapeamento da intencionalidade artificial ..85

4.5 Do impasse computacional aos horizontes da intencionalidade artificial.....103

5 CONCLUSÃO 106

REFERÊNCIAS BIBLIOGRÁFICAS.....109

1. INTRODUÇÃO

Esta dissertação investiga o conceito de *intencionalidade*, definido em seu sentido mais amplo como a propriedade da mente de ser *sobre* algo, de representar ou de se direcionar a objetos, propriedades e estados de coisas (Searle, 1983). A investigação aqui proposta traça a evolução dessa característica, considerada por muitos filósofos como a marca distintiva da mente (Brentano, 1874), desde seus fundamentos filosóficos e suas mais antigas evidências materiais até o desafio contemporâneo imposto pelo desenvolvimento acelerado da inteligência artificial.

O percurso desta pesquisa seguirá uma trajetória histórica e conceitual, iniciando com o problema filosófico da definição da intencionalidade, transitando para a análise das evidências arqueológicas de seu surgimento no período Paleolítico e culminando em um exame de seu estatuto nos modernos sistemas computacionais.

1.1 Motivação da pesquisa: uma trajetória profissional na fronteira da inovação

Uma trajetória profissional de mais de quatro décadas no universo da tecnologia da informação proporcionou o acompanhamento direto das sucessivas revoluções e disrupções que marcaram o setor. Essa jornada permitiu observar a transição de sistemas analógicos para digitais, a ascensão da internet como infraestrutura global, o impacto transformador das redes sociais e, mais recentemente, a evolução da Inteligência Artificial (IA). Cada uma dessas fases desencadeou profundos impactos nos sistemas econômicos, sociais e culturais, revelando a capacidade da tecnologia de reconfigurar a experiência humana em níveis progressivamente mais complexos e interconectados.

A passagem por corporações com um volume significativo de investimento em Pesquisa e Desenvolvimento (P&D), como Microsoft, SAP e Amazon Web Services (AWS), ofereceu uma perspectiva sobre os ciclos de inovação e a pressão competitiva que impulsionam o avanço tecnológico. Contudo, foi em um ambiente acadêmico, durante uma aula de Filosofia da Inteligência Artificial, que as inquietações que motivam esta pesquisa ganharam contornos mais nítidos. A discussão sobre a responsabilização por eventos danosos ou lesivos provocados por

sistemas inteligentes autônomos, tema na obra de pesquisadores como Kate Crawford (2021), despertou uma reflexão intrigante ante aquele debate na audiência, por vezes equilibrado e racional, e por outras inclinado a cenários distópicos. Em minha longa carreira dedicada à tecnologia, a percepção que se consolidou foi a de que as tecnologias não são adversárias, mas parceiras da humanidade, uma visão alinhada à perspectiva do filósofo da técnica Gilbert Simondon. Para Simondon (1980), a cultura deve alcançar um acordo com as entidades técnicas, integrando-as como parte de seu corpo de conhecimento e valores. Essa relação, no entanto, está longe de ser estática ou harmoniosa. O atual estágio da Inteligência Artificial, caracterizado por capacidades como aprendizado profundo (*deep learning*), redes neurais, IA Generativa (*IAGen*) e IA Agêntica (*Agentic* AI), demonstra um progresso no desenvolvimento de sistemas capazes de criar conteúdo e tomar decisões autônomas baseadas em contextos complexos.

Isso levanta uma questão central que motiva esta pesquisa: até que ponto essas tecnologias podem transcender sua funcionalidade programada e adquirir algo comparável à intencionalidade humana? A intencionalidade, definida aqui preliminarmente como a capacidade da mente de se direcionar a ou ser sobre objetos, ideias e ações, é um conceito fundamental tanto na filosofia quanto na compreensão da cognição (Searle, 1983). Ela é considerada uma das marcas distintivas da evolução humana, a faculdade que conecta estados mentais a propósitos e ao mundo. A questão de sua possível emergência em sistemas não biológicos não é, portanto, meramente técnica, mas profundamente filosófica e cultural, tocando no cerne do que significa ser humano. Esta dissertação nasce, assim, da intersecção entre a observação empírica de longo prazo da evolução tecnológica e a inquietação teórica sobre seus limites e destinos.

1.2 O racional da pesquisa: uma jornada em três atos

A presente dissertação estrutura-se como uma jornada de pesquisa exploratória, desenhada para investigar a questão da intencionalidade a partir de uma linha lógica que atravessa três campos fundamentais do conhecimento: a Filosofia da Mente, a Arqueologia e os Sistemas Artificiais. Essa estrutura não é meramente temática, ela constitui uma investigação sobre a própria natureza da causalidade. A pesquisa traça o fenômeno da intencionalidade desde

sua causa biológica e sua história evolutiva material até a possibilidade de sua replicação em um substrato de natureza distinta, o algorítmico. O objetivo é construir um argumento coeso que se desenvolve em três atos, cada um estabelecendo as fundações para o seguinte.

O primeiro ato desta jornada ancora-se na Filosofia da Mente, o campo do saber que investiga a natureza dos fenômenos mentais, como pensamentos, emoções e a própria consciência. Dentro desse campo, a Teoria da Mente, a capacidade de atribuir estados mentais a si e aos outros, torna-se central. Para estabelecer um arcabouço teórico rigoroso, esta dissertação adota a teoria do Naturalismo Biológico do filósofo John Searle. A escolha é metodologicamente crucial. Searle (1983) postula que a intencionalidade é um fenômeno biológico genuíno, causado pelos "poderes causais específicos do cérebro" e irredutível a uma mera simulação computacional. Sua distinção entre a intencionalidade *intrínseca* (original, biológica) e a intencionalidade *derivada* (atribuída a artefatos como textos ou computadores) é o que permite que a questão central desta pesquisa seja formulada de maneira significativa. Tal perspectiva será contrastada com a abordagem funcionalista de Daniel Dennett (1987), para quem a intencionalidade é uma estratégia preditiva útil, a "atitude intencional" (*intentional stance*), aplicada a sistemas complexos para prever seu comportamento. Se a visão de Dennett (1987) fosse adotada, a distinção entre a mente humana e uma "IA sofisticada" seria de grau, não de tipo, dissolvendo o problema que esta dissertação se propõe a investigar. A teoria de Searle (1983), ao contrário, fornece um critério causal e ontológico que torna a investigação possível.

O segundo ato volta-se para as origens, explorando a Arqueologia como uma fonte de evidências materiais da evolução da mente. Define-se Arqueologia como a disciplina que estuda as sociedades humanas do passado através de seus vestígios materiais. Mais especificamente, a Arqueologia Cognitiva busca inferir os processos de pensamento e as capacidades cognitivas de nossos ancestrais a partir de artefatos e de sua distribuição no espaço (Renfrew, 1994). A fabricação de ferramentas de pedra, analisada através do método da *chaîne opératoire* (cadeia operatória), revela a presença de planejamento, antecipação e ação propositada, tornando-se um registro fóssil do comportamento intencional (Galhardo, 2015). A investigação nesse campo é enriquecida por um interesse pessoal profundo, consolidado através da leitura de autores que exploram a pré-história paleolítica; da participação em cursos sobre evolução humana e sobre "Estudo tecnológico de indústrias líticas pré-históricas e pré-coloniais", na Universidade de São

Paulo (USP). Essa base teórica foi complementada por experiências práticas que forneceram uma compreensão mais profunda e "encarnada" do tema. A participação em uma expedição arqueológica em Mostardas (RS – Brasil) e as visitações a sítios arqueológicos de relevância mundial na França – como Lascaux, Les Eyzies, Chauvet-Pont-d'Arc, Gourdon, Grottes Cougnac e Grotte de Domme – permitiram um contato direto com a materialidade da pré-história. Essa vivência prática ofereceu uma camada de conhecimento tácito, um *saber-fazer*, que transcende a análise puramente teórica. Segurar uma ferramenta lítica ou percorrer uma caverna com arte rupestre proporciona uma apreciação das complexidades cognitivas, motoras e sociais envolvidas na sua criação, enriquecendo a compreensão do trabalho de pesquisadores como João Carlos Moreno de Sousa e Sophie de Beaune.

O terceiro e último ato conecta o passado profundo ao futuro emergente, abordando os sistemas artificiais. Inevitável tratar aqui da IA, atualmente posicionada como o mais recente capítulo na longa história da coevolução entre a cognição humana e a tecnologia. É nesse ponto que a questão central da dissertação é formalmente proposta: a intencionalidade permanecerá um atributo irredutível da consciência biológica ou poderá emergir em sistemas artificiais (não biológicos)? O argumento central a ser explorado é que os maciços investimentos, as pesquisas e os desenvolvimentos acelerados em sistemas artificiais pressionam para a emergência de uma forma de intencionalidade intrínseca (original) em substratos não biológicos. Esta jornada tripartida, da filosofia à arqueologia e aos sistemas artificiais não é apenas uma sequência de tópicos, mas a construção de uma linha de base para avaliar uma das questões mais definidoras de nosso tempo.

1.3 Metodologia: fundamentos de uma investigação interdisciplinar

Esta pesquisa adota uma abordagem exploratória, fundamentada em uma revisão bibliográfica e interdisciplinar. O objetivo metodológico é estabelecer uma base teórica para, primeiramente, compreender as origens e a natureza da intencionalidade consciente no ser humano e, em segundo lugar, avaliar criticamente a possibilidade de sua emergência a partir da evolução da Inteligência Artificial.

A revisão bibliográfica foi estruturada em etapas interconectadas para garantir rigor metodológico e profundidade analítica. O processo iniciou-se com um levantamento de fontes

primárias e secundárias, incluindo livros, artigos científicos, teses e documentos relevantes. Esse levantamento foi realizado em um amplo espectro de bases de dados acadêmicas, refletindo a natureza interdisciplinar do estudo: PubMed, SpringerLink, ACM Digital Library, arXiv.org, SciELO, PLOS, Scopus, Web of Science, JSTOR e periódicos especializados como *Journal of Archaeological Science* e *Archaeological Review from Cambridge*. A busca foi orientada por palavras-chave como "intencionalidade", "arqueologia cognitiva", "Inteligência Artificial", "*chaîne opératoire*" e "consciência".

Após a coleta inicial, as fontes foram submetidas a uma triagem baseada em relevância, atualidade e alinhamento com os objetivos da pesquisa. Cada fonte selecionada passou por uma análise crítica para identificar contribuições centrais, argumentos principais, abordagens metodológicas, lacunas e limitações. As informações extraídas foram, então, organizadas em categorias temáticas, estabelecendo as conexões conceituais entre os três campos de investigação da dissertação.

Para garantir a solidez teórica, a pesquisa se apoia em autores-âncora em cada uma de suas áreas de foco:

- No campo da Intencionalidade e da Filosofia da Mente, a discussão é fundamentada nas obras de pensadores clássicos e contemporâneos, como Franz Brentano, John Searle, Edmund Husserl e Daniel Dennett.
- No domínio da Arqueologia, a análise se baseia nos trabalhos de pesquisadores que conectam cultura material e cognição, como Artur Ribeiro, João Carlos Moreno de Sousa, Sophie de Beaune, Collin Refrew, Thomas Wynn, Dietrich Stout, Michael Tomasello e Gilbert Simondon.
- No campo da Inteligência Artificial, a investigação dialoga com os estudos de Roger Penrose, R. R. Poznanski e outros pesquisadores que analisam os impactos, os limites e as possibilidades da emulação da intencionalidade em máquinas.

A síntese final dos dados bibliográficos, integrada às observações provenientes da experiência profissional e do engajamento prático com a arqueologia, visa gerar uma análise interdisciplinar que contribua para o debate sobre a natureza da mente e os horizontes de sua replicação técnica.

1.4 O que é Inteligência Artificial? Desafios de uma definição

Embora este trabalho disserte sobre a Intencionalidade desde suas origens e a possibilidade de sua emergência em sistemas artificiais, trataremos do campo da Inteligência Artificial (IA) nesta sessão, com objetivo de estabelecer seu alicerce conceitual a fim de equipar o leitor com as ferramentas analíticas necessárias para uma compreensão aprofundada dos argumentos sobre Intencionalidade Maquínica, que serão desenvolvidos nos capítulos subsequentes desta dissertação.

O campo da Inteligência Artificial (IA) se desenvolveu a partir de uma questão central: a possibilidade de máquinas manifestarem faculdades tradicionalmente associadas à mente humana. Contudo, a própria tarefa de definir "inteligência" é, em si, complexa e carente de um consenso universal. Essa ambiguidade fundamental moldou a trajetória da IA, que progrediu ao adotar definições pragmáticas e operacionais que permitiram o avanço científico, mesmo na ausência de uma teoria unificada da cognição.

A discussão sobre a natureza da inteligência maquínica é exemplificada pelo contraste entre as visões de proeminentes cientistas. O físico Richard Feynman, por exemplo, postulava que as futuras máquinas não precisariam pensar como os seres humanos para serem eficazes, da mesma forma que um avião não voa como um pássaro; ambos alcançam o mesmo objetivo – voar – por meio de processos, dispositivos e materiais distintos (Kaufman, 2022). Em contrapartida, Marvin Minsky, um dos fundadores do campo, argumentava que os sistemas de IA, embora limitados, já possuíam habilidades de aprendizagem e raciocínio (Kaufman, 2022).

Para organizar as diversas abordagens que surgiram dessa pluralidade de visões, Stuart Russell e Peter Norvig (2021) propuseram um framework taxonômico que classifica os objetivos da IA em duas dimensões: a primeira distingue entre processos de pensamento e comportamento, e a segunda, entre fidelidade ao desempenho humano e adesão a um padrão ideal de racionalidade. Isso resulta em quatro categorias de pesquisa:

Pensar como humanos: a abordagem da modelagem cognitiva, que busca criar programas que simulem o pensamento humano.

Agir como humanos: a abordagem do Teste de Turing, que avalia a capacidade de uma máquina de exibir um comportamento indistinguível do de um ser humano.

Pensar racionalmente: a abordagem das "leis do pensamento", que busca construir sistemas baseados em lógica formal e raciocínio irrefutável.

Agir racionalmente: a abordagem do agente racional, que se concentra na construção de agentes que agem para alcançar o melhor resultado esperado.

Ao longo da história do campo, a abordagem de agir racionalmente consolidou-se como o paradigma dominante. Um agente racional é definido como uma entidade que percebe seu ambiente por meio de sensores e atua sobre ele por meio de atuadores, de forma a maximizar uma medida de desempenho (Russel; Norvig, 2021). Essa definição provou ser mais geral e cientificamente tratável do que as outras, pois o conceito de racionalidade pode ser definido matematicamente e otimizado, permitindo que a IA se desenvolvesse como uma disciplina formal e de engenharia. A evolução do campo pode ser vista como uma trajetória pragmática que se afastou de questões filosóficas insolúveis, como "o que é a consciência?", para se concentrar em um problema de engenharia solucionável: "como construir um sistema que atinge objetivos de forma ótima?".

No entanto, a ênfase no comportamento observável como critério de inteligência não ficou isenta de críticas filosóficas. O argumento do Quarto Chinês, proposto pelo filósofo John Searle, serve como um poderoso contraponto. No experimento mental, Searle se imagina dentro de uma sala fechada, manipulando símbolos chineses com base em um conjunto de regras em inglês (um algoritmo). Para um observador externo, que envia perguntas em chinês e recebe respostas coerentes, pareceria que a sala "entende" chinês. Contudo, Searle, que não compreende chinês, estaria apenas realizando uma manipulação sintática dos símbolos, sem qualquer acesso ao seu significado (semântica) (Searle, 1980, *apud* Penrose, 2023). O argumento sugere que a execução de um algoritmo, por mais complexo que seja, não é uma condição suficiente para a existência de compreensão ou consciência. Essa crítica ataca diretamente a validade do Teste de Turing e reforça a fragilidade das definições de inteligência puramente comportamentais, destacando a distinção crucial entre simular uma ação inteligente e possuir uma mente.

1.5 A genealogia da Inteligência Artificial: uma trajetória histórica

A história da Inteligência Artificial é marcada por uma sucessão de paradigmas, cada um surgindo para superar as limitações do anterior. Essa genealogia revela uma evolução conceitual que se afasta da lógica simbólica e determinística em direção a modelos probabilísticos e generativos, culminando na busca por sistemas autônomos capazes de interagir com o mundo de forma propositiva.

A tabela a seguir sintetiza essa evolução conceitual que será explorada nas sessões a seguir:

Quadro 1 – Evolução dos Paradigmas em Inteligência Artificial

Paradigma	Abordagem central	Exemplo de Tecnologia
IA Simbólica	Manipulação de símbolos com base em regras lógicas explícitas.	Sistemas Especialistas (DENDRAL, MYCIN)
IA Preditiva	Aprendizado de padrões estatísticos e correlações a partir de dados.	Redes Bayesianas, Modelos de Regressão
IA Generativa	Previsão sequencial de tokens para gerar conteúdo novo.	Large Language Models (GPT-4, LLaMA)
IA Agêntica	Integração de modelos para planejar e executar tarefas autônomas.	Agentes de software (Auto-GPT), Robótica Avançada

Fonte: elaborada pelo autor.

1.5.1 A era simbólica: a lógica como fundamento (IA de Dados)

Os primórdios da IA, das décadas de 1950 a 1980, foram dominados pela abordagem simbólica, também conhecida como *Good Old-Fashioned AI* (GOFAI). Esse paradigma era fundamentado na hipótese de que a inteligência poderia ser replicada através da manipulação formal de símbolos com base em regras lógicas explicitamente programadas por especialistas humanos. O conhecimento era codificado em linguagens formais, e os sistemas operavam por meio de inferência lógica para derivar conclusões (Russell; Norvig, 2021).

Os exemplos mais emblemáticos dessa era são os sistemas especialistas. O DENDRAL, desenvolvido na Universidade de Stanford, foi um dos primeiros sucessos, projetado para inferir a estrutura de moléculas a partir de dados de espectrometria de massa. Seu poder derivava não de princípios fundamentais, mas de um vasto conjunto de regras heurísticas (receitas de livro de receitas) extraídas do conhecimento de químicos analíticos (Russell; Norvig, 2021). Outro sistema notável foi o MYCIN, que diagnosticava infecções sanguíneas com base em cerca de 450 regras. Diferentemente do DENDRAL, as regras do MYCIN foram adquiridas por meio de extensas entrevistas com médicos e incorporavam "fatores de certeza" para lidar com a incerteza inerente ao conhecimento médico.

Apesar desses sucessos iniciais, a abordagem simbólica encontrou limitações intransponíveis. A primeira foi a fragilidade: os sistemas eram incapazes de lidar com a incerteza e os dados incompletos do mundo real, que não se encaixavam em suas regras rígidas. A segunda foi a explosão combinatória: a impossibilidade prática de codificar manualmente um número suficiente de regras para cobrir todas as contingências de domínios complexos. Essas dificuldades levaram a promessas não cumpridas e a um subsequente corte de financiamento e interesse, um período que ficou conhecido como o "inverno da IA" (Russell; Norvig, 2021). O fracasso da GOFAI não foi apenas um limite tecnológico, mas um erro epistemológico fundamental. O paradigma presumiu que o conhecimento humano é primariamente explícito e redutível a regras, ignorando o vasto corpo de conhecimento tácito e a capacidade de raciocinar sob incerteza que caracterizam a cognição real.

1.5.2 A revolução probabilística: a ascensão da IA Preditiva

A superação das limitações da IA simbólica veio com uma mudança de paradigma fundamental: em vez de codificar regras manualmente, os sistemas passariam a aprender padrões diretamente dos dados. Essa transição marcou o início da era da IA Preditiva, fundamentada no aprendizado de máquina (*machine learning*) e no raciocínio probabilístico. Essa revolução resolveu o problema da incerteza que assolava a GOFAI, mas ao custo de abandonar a busca por modelos causais e explicativos do mundo (Russell; Norvig, 2021).

O trabalho do cientista da computação Judea Pearl foi seminal para fornecer tanto as ferramentas para essa nova abordagem quanto o arcabouço para compreender suas limitações. Suas redes bayesianas ofereceram um formalismo matemático rigoroso para representar e

raciocinar com a incerteza. Posteriormente, Pearl introduziu a Escada da Causalidade (*Ladder of Causation*), um modelo conceitual que organiza o raciocínio em três níveis cognitivos distintos (Pearl; Mackenzie, 2018):

Associação (ver): o nível mais baixo, que corresponde à capacidade de encontrar padrões e correlações em dados. Responde a perguntas como: "O que um sintoma me diz sobre uma doença?". É o domínio da estatística tradicional.

Intervenção (fazer): o segundo nível, que envolve a predição dos efeitos de ações deliberadas no mundo. Responde a perguntas como: "o que acontecerá se eu tomar este remédio?".

Contrafactuais (imaginar): o nível mais alto, que representa a capacidade de imaginar mundos alternativos e raciocinar sobre causas e efeitos em retrospecto. Responde a perguntas como: "Teria minha dor de cabeça passado se eu não tivesse tomado o remédio?".

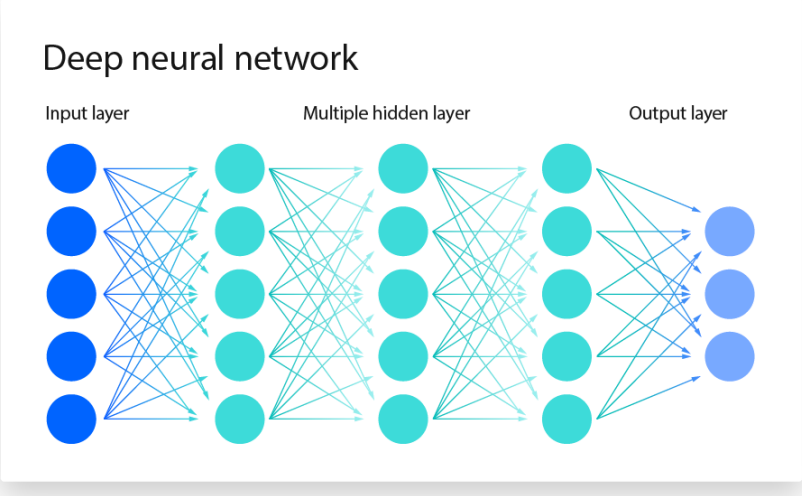
A IA Preditiva, incluindo os mais avançados modelos de *deep learning* (aprendizado profundo), opera quase exclusivamente no primeiro degrau da escada, o da associação (Pearl; Mackenzie, 2018). Esses sistemas são extremamente eficazes em detectar que fumaça está correlacionada com fogo, mas não compreendem a relação causal de que o fogo causa a fumaça. Seu "conhecimento" é derivado de padrões estatísticos em dados que foram coletados e rotulados por agentes humanos. Portanto, o sucesso da IA moderna é o sucesso da correlação em escala massiva. A Escada da Causalidade revela que, apesar de sua complexidade, esses modelos são fundamentalmente cegos para a estrutura causal do mundo, uma limitação inerente que os impede de realizar um planejamento robusto ou raciocinar sobre as consequências de ações inéditas.

1.5.3. O motor da previsão: redes neurais profundas (*deep learning*)

A arquitetura computacional que tornou a revolução probabilística uma realidade prática é a rede neural profunda (*deep neural network*), e a técnica associada é conhecida como aprendizado profundo (*deep learning*). Inspirados na estrutura do cérebro biológico, esses modelos consistem em unidades de processamento simples (os "neurônios artificiais") organizadas em múltiplas camadas interconectadas (Cozman; Kaufman, 2022). A informação flui da camada de entrada (*input*), que recebe os dados brutos (como os pixels de uma imagem),

através de uma série de "camadas ocultas", até uma camada de saída (*output*), que produz a previsão final (Cozman; Kaufman, 2022).

Figura 1 – Deep Neural Network



Fonte: IBM (Disponível em: https://www.ibm.com/br-pt/think/topics/neural-networks). A imagem apresentada ilustra a arquitetura típica de uma *Deep Neural Network* (DNN), uma classe de redes neurais artificiais caracterizada pela presença de múltiplas camadas ocultas (*hidden layers*) entre a camada de entrada (*input layer*) e a camada de saída (*output layer*).

O adjetivo "profunda" refere-se precisamente à presença de múltiplas camadas ocultas. Cada camada aprende a detectar padrões em um nível de abstração diferente. Por exemplo, em uma tarefa de reconhecimento de imagem, a primeira camada pode aprender a reconhecer arestas e cores simples. A segunda camada pode combinar essas arestas para identificar formas mais complexas, como olhos ou narizes. As camadas subsequentes combinam essas formas para reconhecer objetos inteiros, como rostos ou carros (Cozman; Kaufman, 2022). Essa hierarquia de abstração permite que os modelos de *deep learning* aprendam padrões extremamente complexos e sutis a partir de dados de alta dimensionalidade, superando em muito as capacidades de modelos de aprendizado de máquina mais rasos.

O aprendizado ocorre durante a fase de treinamento, na qual a rede é alimentada com um vasto conjunto de dados rotulados (por exemplo, milhões de imagens, cada uma com um rótulo indicando se é um "gato" ou um "cachorro"). Inicialmente, as conexões entre os neurônios (os

"pesos") são aleatórias. Para cada dado de entrada, a rede faz uma previsão. Essa previsão é comparada com o rótulo correto, e um "erro" é calculado. O algoritmo de retropropagação (*back-propagation*), então, ajusta ligeiramente todos os pesos da rede de uma maneira que reduza esse erro. Repetido milhões de vezes, esse processo faz com que os pesos da rede convirjam para uma configuração que captura com precisão os padrões estatísticos presentes nos dados de treinamento (Cozman; Kaufman, 2022). O momento decisivo para essa abordagem ocorreu em 2012, quando uma rede neural profunda venceu a competição de reconhecimento de imagem ImageNet com uma margem de erro drasticamente menor que a dos concorrentes, demonstrando a superioridade da técnica e inaugurando a era moderna da IA (Kaufman, 2022).

1.5.4 A fronteira criativa: o advento da IA Generativa

A evolução mais recente do paradigma de aprendizado de máquina deu origem à IA Generativa. Esse subcampo possui um foco distinto: em vez de apenas analisar ou classificar dados existentes, seu objetivo é gerar conteúdo inteiramente novo e complexo, como textos, imagens, áudios e códigos de programação. A distinção conceitual é crucial: a IA Preditiva recebe dados como entrada para produzir uma previsão ou classificação (por exemplo, analisar uma imagem para classificá-la como "gato"), enquanto a IA Generativa utiliza padrões aprendidos em dados para sintetizar um novo artefato (por exemplo, gerar a imagem de um gato que nunca existiu) (Santaella; Kaufman, 2024).

Esses modelos têm o potencial de impactar significativamente a economia criativa, com aplicações que vão desde a geração automatizada de conteúdo de marketing até a produção de peças publicitárias e artigos científicos. A diversidade de modelos generativos pode ser organizada em uma taxonomia que ilustra a amplitude de suas capacidades.

Quadro 2 – Diversidade de modelos generativos

Categoria	Descrição	Modelos Ilustrativos
Text-to-Image	Gera imagens a partir de descrições textuais (*prompts*).	DALL-E 2, Stable Diffusion, Midjourney

Text-to-Text	Gera texto novo e coerente a partir de um *prompt* textual.	ChatGPT (série GPT), LaMDA
Text-to-Video	Gera sequências de vídeo a partir de descrições textuais.	Phenaki, Soundify
Text-to-Code	Gera código de programação em diversas linguagens.	Codex, AlphaCode
Outras	Text-to-Audio, Image-to-Text, Text-to-3D etc.	Jukebox, Flamingo, Dreamfusion

Fonte: Santaella; Kaufman, 2024.

A emergência da IA Generativa representa um salto na capacidade dos sistemas de IA de manipular e recombinar informações de maneiras que imitam a criatividade humana, embora os mecanismos subjacentes permaneçam fundamentalmente estatísticos.

1.6 A mecânica dos modelos multifuncionais: a arquitetura *transformer*

O avanço exponencial da IA Generativa, especialmente em tarefas de linguagem, foi impulsionado por uma inovação arquitetônica fundamental: o Transformer. Proposta em 2017 por pesquisadores do Google, essa arquitetura superou as limitações de modelos anteriores e se tornou a base para os grandes modelos de linguagem (*Large Language Models*, LLMs), como a série GPT que alimenta o ChatGPT (Santaella; Kaufman, 2024).

O componente central do Transformer é o mecanismo de autoatenção (*self-attention*). Modelos anteriores, como as Redes Neurais Recorrentes (RNNs), processavam o texto sequencialmente, palavra por palavra. Isso criava um gargalo informacional, dificultando a captura de relações de longo alcance em um texto; o modelo tendia a "esquecer" o contexto de palavras que apareceram muito antes. O mecanismo de autoatenção resolve esse problema ao processar todas as palavras de uma sequência de entrada simultaneamente. Para cada palavra, ele calcula um "peso de atenção" que mede a relevância de todas as outras palavras na sequência para a compreensão do significado daquela palavra específica. Isso permite que o modelo construa uma representação contextual rica, capturando dependências complexas independentemente da distância entre as palavras (Vaswani *et al*., 2017).

Para compreender como esses modelos operam, é essencial distinguir duas fases operacionais distintas: treinamento e inferência.

Treinamento (aprendizado): fase de pré-computação, extremamente intensiva em recursos computacionais e tempo. O processo inicia com um modelo contendo milhões ou bilhões de parâmetros (também chamados de pesos) com valores aleatórios. Esse modelo é então alimentado com um vasto *corpus* de dados, como uma grande fração da internet. A tarefa de treinamento consiste em, repetidamente, prever a próxima palavra em uma sequência de texto. A cada previsão, o modelo compara sua saída com a palavra correta e calcula um erro. O algoritmo de retropropagação (*back-propagation*) utiliza esse erro para ajustar ligeiramente todos os parâmetros do modelo na direção que minimiza o erro futuro. Esse ciclo de previsão, cálculo de erro e ajuste de pesos é repetido trilhões de vezes, um processo que pode levar semanas ou meses em supercomputadores. Ao final, os pesos do modelo convergem para valores que capturam os padrões estatísticos, sintáticos e semânticos da linguagem presente nos dados de treinamento (Cozman; Kaufman, 2022).

Inferência (geração): é a fase de uso do modelo, um processo comparativamente rápido e computacionalmente leve. O modelo treinado, com seus parâmetros agora fixos, recebe uma nova entrada (um *prompt*) de um usuário. Utilizando os padrões consolidados durante o treinamento, ele calcula a distribuição de probabilidade para a próxima palavra mais provável. Essa palavra é gerada e anexada à sequência, e o processo se repete, gerando a resposta palavra por palavra. Nessa fase, não ocorre mais aprendizado; o modelo está apenas aplicando o conhecimento pré-compilado.

A separação entre treinamento e inferência é a chave para entender tanto a capacidade quanto as limitações da IA Generativa. Sua "inteligência" e fluidez aparentes são, na verdade, a aplicação em alta velocidade de padrões estatísticos aprendidos de um conjunto de dados estático e congelado no tempo. Isso explica por que os modelos não têm conhecimento de eventos ocorridos após a data de corte de seu treinamento e por que podem "alucinar" – inventar fatos ou fontes – quando um *prompt* os leva a uma região do espaço de probabilidade onde os dados de treinamento eram esparsos. Eles não possuem um modelo de mundo, uma noção de verdade ou capacidade de raciocínio em tempo real; apenas uma capacidade extraordinariamente sofisticada de prever a próxima palavra mais provável.

1.7 Rumo à autonomia: a emergência das IAs Agênticas

Enquanto a IA Preditiva e a Generativa são modelos passivos que respondem a entradas, a fronteira da pesquisa em IA avança para um paradigma mais ativo e integrado: a IA Agêntica. Esse conceito não se refere a um novo tipo de algoritmo de aprendizado, mas a uma abordagem arquitetônica que combina diversas capacidades de IA para criar sistemas que percebem, raciocinam e atuam de forma autônoma para atingir objetivos em ambientes complexos.

A base conceitual para a IA Agêntica é o paradigma do agente racional, conforme definido por Russell e Norvig (2021). Um agente é um sistema que interage com um ambiente: ele o percebe através de sensores e age sobre ele através de atuadores. A racionalidade de um agente é medida por sua capacidade de selecionar ações que maximizem uma medida de desempenho esperada, dadas suas percepções e seu conhecimento. A complexidade e a capacidade de um agente podem ser descritas em uma hierarquia:

Agentes reativos simples: operam com base em regras de condição-ação ("se o sensor de sujeira detectar sujeira, então ligue a sucção"). Eles não mantêm memória do passado.

Agentes baseados em modelo: mantêm um estado interno que representa sua crença sobre o estado atual do mundo. Isso lhes permite lidar com ambientes parcialmente observáveis, inferindo aspectos não vistos com base no histórico de percepções e em um modelo de como o mundo funciona.

Agentes baseados em objetivos: além de um modelo do mundo, possuem uma descrição de estados desejáveis (objetivos). Suas decisões envolvem busca e planejamento para encontrar sequências de ações que levem a esses objetivos.

Agentes baseados em utilidade: quando múltiplos objetivos existem ou quando há incerteza sobre os resultados das ações, uma função de utilidade quantifica a desejabilidade de diferentes estados do mundo. O agente age para maximizar sua utilidade esperada, permitindo-lhe lidar com objetivos conflitantes e tomar decisões sob risco.

Uma IA Agêntica moderna representa a instanciação de uma dessas arquiteturas – tipicamente as mais sofisticadas, baseadas em modelo, objetivos e utilidade – através da integração de múltiplos modelos de IA especializados. Por exemplo, um robô doméstico

autônomo poderia usar um grande modelo de linguagem para compreender comandos verbais, um modelo de visão computacional para perceber objetos e navegar no ambiente, e um modelo de aprendizado por reforço para aprender a executar tarefas físicas.

Essa integração representa uma tentativa de superar a principal limitação da IA Preditiva e Generativa: sua passividade e sua incapacidade de raciocínio causal. Um agente, para agir eficazmente no mundo, não pode depender apenas de correlações passadas. Ele precisa de um modelo causal interno, mesmo que implícito, que lhe permita subir na Escada da Causalidade de Pearl. A ação autônoma exige, no mínimo, a capacidade de raciocinar no segundo degrau: o da intervenção. O agente deve ser capaz de se perguntar: "O que acontecerá se eu executar a ação A em vez da ação B?". Essa capacidade de prever as consequências das próprias ações é a essência do planejamento e da tomada de decisão inteligente.

Portanto, a IA Agêntica não é apenas a próxima etapa cronológica na evolução da IA – é uma síntese necessária que força o campo a confrontar o problema da causalidade, que foi amplamente contornado pela revolução probabilística. A verdadeira autonomia não emerge da simples identificação de padrões, mas da capacidade de modelar o mundo e intervir nele de forma propositiva. Isso fecha o arco narrativo da evolução da IA, mostrando um retorno a problemas fundamentais de representação do conhecimento e raciocínio, agora equipados com as ferramentas estatísticas e computacionais da era moderna.

2. A INTENCIONALIDADE

Este capítulo estabelece o fundamento conceitual da presente dissertação, abordando a definição de Intencionalidade. A questão central que norteia esta dissertação – se a intencionalidade, a propriedade da mente de ser "sobre" ou "direcionada a" objetos e estados de coisas, permanecerá um atributo irredutível da consciência biológica ou poderá emergir em sistemas artificiais não biológicos – exige um arcabouço teórico que não apenas defina o fenômeno, mas que também forneça critérios robustos para a sua identificação. A teoria de Searle (2006) oferece a estrutura coerente e empiricamente fundamentada para esta investigação. A sua força reside na postulação da intencionalidade como um fenômeno biológico genuíno, causado pelos poderes causais específicos do cérebro e intrinsecamente ligado à ontologia da consciência (Searle, 2006, 1990).

O Naturalismo Biológico de Searle, ao contrário de abordagens funcionalistas ou eliminativistas, preserva a realidade ontológica e a eficácia causal dos estados mentais subjetivos, fornecendo um critério claro – embora neurobiologicamente ainda não especificado em detalhe – para distinguir a intencionalidade genuína (intrínseca) de suas simulações (derivada ou "como-se"). Essa distinção é indispensável para a análise da intencionalidade na pré-história, que será o foco do Capítulo 3, e para a investigação de sua possível emergência em sistemas algorítmicos, tema do Capítulo 4. A escolha por Searle não é meramente uma preferência filosófica, mas uma decisão metodológica fundamental para a totalidade da dissertação. Ao definir a intencionalidade como um produto de "características causais do cérebro", Searle (2021, p. 1) estabelece um critério potencialmente falsificável para a IA: um sistema artificial só possuirá intencionalidade se puder replicar esses poderes causais específicos, e não apenas emular a sua função ou comportamento externo. Essa formulação transforma uma questão tradicionalmente filosófica em um problema que se abre à investigação empírica da neurociência e da engenharia de IA. A questão deixa de ser apenas sobre o que um sistema *faz* (seu *output*) e passa a ser sobre *como* e *com que substrato* ele o faz, ou seja, a maquinaria causal subjacente. Esse critério de demarcação causal-biológico é o que permite uma análise coerente tanto do passado evolutivo da mente humana quanto do futuro potencial da mente artificial.

2.1. O Naturalismo Biológico: a intencionalidade como fenômeno causal do cérebro

O Naturalismo Biológico, conforme formulado por Searle, postula que os fenômenos mentais, incluindo a consciência e a intencionalidade, são características biológicas de nível superior de determinados organismos, especificamente do cérebro e do sistema nervoso central (Searle, 2006, p. 127; Lyra; Mograbi; El-Hani, 2016). Essa abordagem busca resolver o tradicional problema mente-corpo evitando tanto as armadilhas do dualismo cartesiano, que postula duas substâncias distintas e ontologicamente separadas (mental e física), quanto as do materialismo reducionista, que frequentemente nega a realidade ou a eficácia causal dos estados mentais subjetivos.

Para Searle (2006), os estados mentais são, simultaneamente, causados por processos neurofisiológicos de nível inferior (como os padrões de disparos de neurônios em sinapses) e realizados na estrutura do cérebro como um sistema completo (Searle, 2006, p. 180). A relação não é de identidade redutiva, mas de emergência causal. Searle utiliza a analogia com propriedades físicas da matéria para elucidar esse ponto: a liquidez da água, por exemplo, é uma propriedade de nível superior (macro) que é *causada por* interações no nível inferior (micro) do comportamento das moléculas de H_2O. A liquidez é, ao mesmo tempo, uma característica *realizada no* sistema de moléculas; uma molécula individual não é líquida, mas o sistema de moléculas, sob certas condições, é (Searle, 2006). De forma análoga, a consciência ou um estado intencional como a sede não se encontram em um neurônio individual, mas são características de nível superior do sistema cerebral como um todo, causadas pela atividade neuronal coletiva (Searle, 2006).

Desta forma, a mente não é uma substância imaterial pairando sobre o cérebro, nem é "nada além de" a soma de seus componentes neuronais. É uma propriedade biológica emergente, tão real e causalmente eficaz quanto a digestão ou a fotossíntese (Searle, 2002). Essa perspectiva naturalista integra a mente na ordem natural, tratando-a como um objeto de estudo científico legítimo, sem despojá-la de suas características essenciais, como a subjetividade e a intencionalidade.

2.2. Intencionalidade intrínseca vs. derivada e "como-se"

Uma consequência direta do Naturalismo Biológico é a necessidade de distinguir entre diferentes tipos de "direcionalidade" ou "sobreidade" que encontramos no mundo. Searle (2006) propõe uma taxonomia tripartite que se revela fundamental para a análise da mente e para a avaliação das pretensões da Inteligência Artificial.

1. *Intencionalidade intrínseca*: é o fenômeno genuíno e original da intencionalidade. É uma propriedade que seres humanos e certos outros animais possuem como parte de sua natureza biológica. Não é uma questão de interpretação externa, mas um fato sobre a constituição do organismo. Quando um animal sente sede, teme um predador ou vê um objeto, estes são estados intencionais intrínsecos, causados por sua neurofisiologia

(Searle, 2002; 2006). É a "coisa real" que a filosofia da mente e a ciência cognitiva devem explicar.

2. *Intencionalidade derivada*: essa forma de intencionalidade é real, mas não é intrínseca ao objeto que a possui. Ela é, como o nome sugere, derivada da intencionalidade intrínseca de agentes mentais, como os seres humanos. O exemplo paradigmático é o significado linguístico. A frase em francês "j'ai grand soif" significa "estou com muita sede" (Searle, 2006, p. 118). A sequência de marcas de tinta ou de sons não possui essa direcionalidade por si só; ela a adquire porque uma comunidade de falantes com intencionalidade intrínseca (que sentem sede de verdade) a utiliza para expressar e comunicar seus estados mentais. Mapas, diagramas e retratos são outros exemplos de objetos com intencionalidade derivada (Searle, 2006, p. 118).
3. Intencionalidade "Como-se" (*As-if*): essa não é uma forma de intencionalidade, mas uma atribuição metafórica ou figurativa. Descrevemos sistemas que se comportam *como se* tivessem intencionalidade. Dizer "Meu gramado está com sede" ou "O termostato 'sabe' quando a sala está fria" é usar a linguagem de forma instrumental (Searle, 2006, p. 118). Tais atribuições são psicologicamente irrelevantes, pois não implicam a presença de nenhum fenômeno mental real no gramado ou no termostato. Searle (2006) adverte que rejeitar essa distinção leva ao panpsiquismo, a visão de que tudo no universo é mental. Se uma pedra caindo "deseja" chegar ao centro da Terra, então a noção de mentalidade perde todo o seu poder explicativo).

Essa taxonomia tripartite é a ferramenta analítica que permite à presente dissertação abordar a questão da IA de forma rigorosa. Um programa de computador, por mais sofisticado que seja, se enquadra, na visão de Searle, nas categorias de intencionalidade derivada (seus símbolos e algoritmos recebem significado de seus programadores) e "como-se" (seu comportamento externo simula compreensão ou pensamento). A partir desse saber, a questão central da dissertação pode ser reformulada com maior precisão: pode um sistema artificial, através de sua própria operação, transcender a intencionalidade derivada e "como-se" para gerar intencionalidade intrínseca?

2.3. A estrutura da intencionalidade intrínseca

Para compreender plenamente a intencionalidade intrínseca, é necessário analisar sua estrutura interna e os mecanismos que a sustentam. Segundo Searle (2002), não se trata de uma propriedade monolítica, mas de um fenômeno complexo com componentes, mecanismos e estruturas de suporte definidos.

2.3.1. Componentes essenciais e mecanismos de funcionamento

Todo estado intencional é composto por dois elementos inseparáveis: um *conteúdo intencional* e um *modo psicológico* (Searle, 2002; Canal, 2006). O *conteúdo* é a representação em si (e.g., *que está chovendo*, *que eu erga meu braço*), enquanto o *modo* é a atitude que a mente assume em relação a esse conteúdo (e.g., *crença*, *desejo*, *intenção*, *medo*). Não se pode ter uma crença sem um conteúdo, nem um conteúdo proposicional flutuando livremente sem um modo psicológico que o sustente.

A partir dessa estrutura, emergem os mecanismos de funcionamento da intencionalidade:

- *Condições de satisfação:* o conteúdo intencional determina as condições que devem se dar no mundo para que o estado seja "satisfeito". Uma crença é satisfeita se for verdadeira; um desejo é satisfeito se for realizado; uma intenção é satisfeita se for executada. Esse é o critério pelo qual a relação entre mente e mundo é avaliada (Searle, 2002; Canal, 2006).
- *Direção de ajuste*: esse conceito crucial descreve a relação de responsabilidade entre a mente e o mundo para que as condições de satisfação sejam alcançadas. Searle (2002) identifica duas direções:
    - *Direção de ajuste mente-mundo (mind-to-world):* a mente é responsável por se ajustar à realidade. Esse é o caso de estados cognitivos como crenças e percepções. Se minha crença de que está chovendo não corresponde ao fato de que o sol está brilhando, é a minha crença que está em falha e deve ser alterada, não o mundo.
    - *Direção de ajuste mundo-mente (world-to-mind):* o mundo é responsável por se ajustar à mente. Esse é o caso de estados volitivos, como desejos e intenções. Se eu desejo que chova, o desejo só será satisfeito se o mundo mudar para se conformar ao meu estado mental. A falha está no mundo por não realizar meu desejo.

- *Causalidade autorreferencial:* certos estados intencionais, notadamente a percepção e a ação, possuem uma condição de satisfação adicional e interna: eles devem estar na relação causal correta com seu objeto. Minha experiência visual de um carro só conta como uma percepção genuína se for *causada pela presença e pelas características daquele carro*. Da mesma forma, minha intenção de erguer o braço só é satisfeita se *essa mesma intenção causar* o movimento do braço. Uma cadeia causal desviante (e.g., minha intenção me deixa nervoso, o nervosismo causa um espasmo que ergue meu braço) não satisfaz a intenção original (Searle, 2002)

2.4. Estruturas de suporte: a rede e o *background*

Os estados intencionais não operam em um vácuo. Searle (2002) argumenta que eles dependem de duas estruturas de suporte interligadas: a rede e o *background* A rede *(the network*): nenhum estado intencional, como uma crença ou um desejo, funciona de forma isolada ou atomística. Ele só tem sentido e pode operar dentro de uma vasta rede holística de outros estados intencionais. Para ter o desejo de pedir uma refeição em um restaurante, é preciso ter uma multitude de outras crenças e desejos: a crença de que restaurantes existem, de que eles servem comida mediante pagamento, de que a comida aliviará a fome, o desejo de comer etc. (Searle, 2002; 2006).

- O *background*: mais fundamentalmente, a própria rede só funciona sobre um *background* de capacidades, habilidades, posturas e saberes-fazer que são, em si, pré-intencionais e não representacionais (Searle, 2002; 2006; Canal, 2006). Não temos uma "crença" explícita de que o chão sob nossos pés é sólido ou que objetos não desaparecem quando não os olhamos – simplesmente *damos isso por certo* em nosso comportamento e em nossas interações com o mundo. Esse *background* de práticas corporificadas e pressuposições não representadas é o que permite a interpretação e a aplicação de nossos estados intencionais. Sem ele, o conteúdo intencional seria radicalmente indeterminado. Ao pedir um bife em um restaurante, não precisamos especificar que ele não deve ser entregue em nossa casa, ou que não deve vir envolto em concreto, ou que não deve ser esparramado em nossa cabeça (Searle, 2006). Essas infinitas possibilidades são pré-reflexivamente excluídas pelo *background*.

O conceito de *background* representa um dos desafios mais profundos para a Inteligência Artificial simbólica tradicional (GOFAI). Enquanto a IA, na sua vertente clássica, tal como criticada por filósofos como Hubert Dreyfus (1972), tentou modelar a inteligência através da manipulação de representações simbólicas (regras e proposições explícitas, que corresponderiam à rede de Searle), o argumento de Searle (2006) sobre o *background* postula que a representação só funciona porque está ancorada em um vasto conjunto de capacidades não representacionais e corporificadas. Um sistema de IA que tentasse codificar explicitamente todas as pressuposições do *background* enfrentaria uma explosão combinatória intratável, um problema intimamente relacionado ao "*frame problem*". Portanto, segundo Searle (2006), o *background* não é um mero detalhe filosófico, é uma objeção fundamental ao projeto de criar intencionalidade através da pura manipulação de símbolos, apontando para a necessidade de uma abordagem que leve em conta a corporificação e a interação com o mundo, o que será de importância crucial para a análise no capítulo 4.

2.5. A conexão indissociável: intencionalidade, consciência e cognição

Para Searle, qualquer teoria da mente que negligencie ou marginalize o fenômeno da consciência comete um erro fundamental. A consciência não é apenas mais uma propriedade da mente, ela é o centro a partir do qual todos os outros fenômenos mentais significativos, incluindo a intencionalidade, devem ser compreendidos (Searle, 2006; Lyra; Mobrabi; El-Hani, 2016).

2.5.1. O princípio da conexão e a natureza do inconsciente

A tese mais forte de Searle (1990) sobre essa relação é o *Princípio da Conexão*, que afirma que "todos os estados intencionais inconscientes são, em princípio, acessíveis à consciência" (p. 585). Não existe, para Searle, um domínio de intencionalidade intrínseca que seja profundamente e em princípio inacessível. Essa afirmação tem consequências radicais para a ciência cognitiva.

Para fundamentar esse princípio, Searle estabelece uma distinção ontológica crucial entre o que é *inconsciente* e o que é *não consciente*:

- *Estados mentais inconscientes (unconscious*): são estados mentais genuínos que, em um dado momento, não estão no foco da atenção consciente. Exemplos incluem a maioria de nossas crenças (como a crença de que a Torre Eiffel está em Paris, que mantemos mesmo durante o sono); memórias não evocadas ou desejos reprimidos. O que torna essas configurações neurofisiológicas estados *mentais*, e não apenas fatos neurológicos brutos, é a sua capacidade causal de produzir o estado consciente correspondente. A ontologia de um estado mental inconsciente é, portanto, a de uma disposição neurofisiológica para causar consciência. Mesmo que essa capacidade seja bloqueada (por exemplo, por lesão cerebral ou repressão), o estado continua sendo do *tipo* de coisa que poderia ter sido consciente (Searle, 1990; 2006).
- *Processos não conscientes (nonconscious*): são processos puramente neurofisiológicos que são essenciais para a vida mental, mas que não são, eles próprios, estados mentais e não são, em princípio, acessíveis à consciência. A mielinização dos axônios, as secreções hormonais que regulam o humor, ou o funcionamento do reflexo vestíbulo-ocular (ROV) são exemplos de fenômenos não conscientes (Searle, 1990). Eles são parte da maquinaria causal que suporta a mente, mas não são, eles mesmos, mentais no sentido intencional.

A consciência, por sua vez, possui uma *ontologia de primeira pessoa*, ou subjetividade, que é irredutível (Searle, 2006). A experiência da dor, por exemplo, tem um modo de existência que é "sentir-se como algo". Esse aspecto qualitativo e subjetivo não pode ser plenamente capturado por uma descrição objetiva, de terceira pessoa, como um padrão de atividade neuronal. Negar essa ontologia de primeira pessoa é, para Searle, negar o próprio fenômeno que se pretende explicar.

2.5.2 A inversão explicativa e suas implicações metodológicas

O *Princípio da Conexão* e a distinção entre inconsciente e não consciente levam diretamente ao que Searle chama de *Inversão Explicativa*, uma ferramenta metodológica de grande alcance (Searle, 1990). Muitas teorias na ciência cognitiva postulam regras mentais profundamente inconscientes para explicar o comportamento (e.g., "Seguimos a regra X para formar frases gramaticais"). Essa é uma explicação intencionalista: um conteúdo mental (a regra) causa o comportamento.

O Princípio da Conexão de Searle proíbe a existência de tais regras como fenômenos mentais intrínsecos se elas forem, em princípio, inacessíveis à consciência. Tais processos devem ser reclassificados como *não conscientes* (Searle, 1990). Consequentemente, a estrutura da explicação deve ser invertida, seguindo um modelo análogo ao que a Teoria da Evolução de Darwin impôs às explicações teleológicas pré-darwinianas.

A explicação invertida procede em dois níveis:

1. *Nível causal (Hardware):* um mecanismo neurofisiológico (ou, por analogia, computacional) *causa* um padrão de comportamento. A causa é mecânica, não intencional.
2. *Nível funcional (teleológico):* esse padrão de comportamento tem uma *função* adaptativa ou útil para o organismo em seu ambiente.

Não é a planta que "deseja" a luz solar e, portanto, "vira" suas folhas (explicação intencionalista). Em vez disso, secreções hormonais (auxina) *causam* o movimento das folhas, e esse movimento tem a *função* de maximizar a fotossíntese, o que confere uma vantagem de sobrevivência (Searle, 1990).

Essa inversão tem implicações diretas para as investigações desta dissertação. Na análise arqueológica do capítulo 3, ao observar um padrão regular no lascamento de ferramentas de pedra, a inversão nos impede de saltar para a conclusão de que "o hominínio seguia uma regra mental". Em vez disso, nos leva a perguntar: (a) quais mecanismos neuromotores e capacidades cognitivas *causaram* esse padrão repetitivo?; e (b) qual a *função* adaptativa dessa forma de ferramenta? Similarmente, na análise da IA no capítulo 4, em vez de dizer que "o programa segue uma regra para entender chinês", a inversão nos força a reconhecer que (i) processos computacionais puramente sintáticos *causam* a manipulação de símbolos, e (ii) essa manipulação tem a *função* de produzir um *output* que *nós* interpretamos como compreensivo. A inversão expõe que a IA, em sua forma atual, opera no nível do mecanismo não consciente, carecendo da intencionalidade intrínseca que só poderia advir da capacidade de tornar a "regra" consciente.

2.6 A perspectiva evolutiva: a origem biológica e os níveis da intencionalidade

A teoria de Searle (2006) posiciona a intencionalidade e a consciência não como mistérios filosóficos ou propriedades transcendentais, mas como produtos concretos da evolução biológica. Elas são parte da história natural da vida, tão integradas à biologia quanto a digestão, a mitose ou a fotossíntese. Essa perspectiva é fundamental para entender tanto a sua natureza quanto a sua função.

2.6.1 A vantagem seletiva da consciência e da intencionalidade

A consciência, para Searle (2006), não é um epifenômeno – um mero subproduto sem função causal. Pelo contrário, ela confere uma vantagem seletiva crucial. O autor ilustra esse ponto analisando casos de comportamento complexo em pacientes inconscientes (e.g., durante crises de epilepsia *petit mal*). Tais pacientes podem executar tarefas rotineiras e memorizadas, como dirigir um carro ou tocar piano, de forma automática. Isso demonstra que mecanismos não conscientes podem sustentar comportamentos complexos. No entanto, esses comportamentos carecem em flexibilidade, sensibilidade ao contexto e criatividade. A consciência, argumenta Searle, é o que adiciona precisamente essas capacidades. Ela permite uma gama muito maior de discriminações e uma resposta muito mais adaptável e criativa aos desafios do ambiente, o que representa uma clara vantagem evolutiva

A normatividade inerente à intencionalidade – noções como “verdade”, “falsidade”, “sucesso” e “falha” – também encontra seu fundamento na biologia evolutiva. A racionalidade não é imposta por uma lógica abstrata, mas é o resultado de pressões seletivas. Um organismo cujos sistemas perceptivos e de crenças não correspondem de forma confiável ao seu ambiente (ou seja, não produzem representações “verdadeiras” ou “bem-sucedidas”) tem uma probabilidade drasticamente menor de sobreviver e se reproduzir. Assim, a “lógica” da intencionalidade está profundamente enraizada na “lógica” da sobrevivência, ancorando firmemente a filosofia da mente na biologia (Searle, 1990).

2.6.2 Níveis de intencionalidade em outros seres vivos

A abordagem de Searle (2021) permite uma análise nuançada e respeitosa dos diferentes níveis de intencionalidade no reino animal. A atribuição de intencionalidade a outros seres não se baseia apenas em uma analogia comportamental, mas em uma presunção informada sobre uma base causal fisiológica semelhante à nossa.

Reconhecemos com confiança a intencionalidade em mamíferos como cães e primatas porque, além de seu comportamento complexo e flexível, eles possuem sistemas nervosos, cérebros e órgãos dos sentidos que são anatomicamente e fisiologicamente análogos aos nossos. Supomos, razoavelmente, que essas estruturas homólogas possuem poderes causais semelhantes para produzir experiências conscientes e estados intencionais (Searle, 2006; 2021). A atribuição não é uma mera projeção sentimental, mas uma inferência baseada na continuidade biológica.

Para organismos mais distantes na árvore filogenética, como insetos (e.g., gafanhotos, pulgas), a atribuição torna-se mais incerta e constitui um caso limítrofe (Searle, 2021). A questão de saber se tais criaturas possuem intencionalidade intrínseca torna-se um problema empírico para a neurobiologia comparada, não uma questão a ser decidida *a priori*. Essa abordagem evita um antropocentrismo simplista e trata a intencionalidade como um contínuo biológico, com diferentes graus de complexidade e sofisticação que dependem diretamente da estrutura neurofisiológica subjacente de cada organismo. A análise é, portanto, sempre conduzida com respeito às diferentes formas de vida, evitando hierarquias de valor e baseando-se em evidências biológicas.

2.6.3 A justificativa da escolha teórica: Searle em contraponto a outras perspectivas

A adoção do Naturalismo Biológico como o arcabouço teórico desta dissertação requer uma justificativa robusta, especialmente quando confrontada com alternativas influentes. A força da posição de Searle torna-se mais evidente ao ser contrastada com a visão da Inteligência Artificial Forte e, principalmente, com a abordagem funcionalista de Daniel Dennett (1987).

2.7. O desafio da Inteligência Artificial forte e o Argumento do Quarto Chinês

A Inteligência Artificial Forte (IA Forte) é a tese de que a mente é para o cérebro o que o software é para o hardware. Ela sustenta que instanciar um programa de computador com a complexidade e os *inputs/outputs* corretos é, por si só, uma condição suficiente para a existência de estados mentais e intencionalidade (Searle, 2021).

Para refutar essa tese, Searle (2021) desenvolveu seu célebre experimento mental, o Argumento do Quarto Chinês. Nele, um indivíduo que não entende uma palavra de chinês é trancado em um quarto e recebe um conjunto de regras em inglês (o "programa") que lhe permite manipular símbolos chineses. Ele recebe "perguntas" em chinês e, seguindo as regras, produz "respostas" em chinês que são indistinguíveis das de um falante nativo. Do ponto de vista externo, o sistema (homem + regras) passa no Teste de Turing para a compreensão do chinês. No entanto, o homem no quarto continua sem entender absolutamente nada de chinês.

A conclusão central do argumento é que a sintaxe não é suficiente para a semântica. O homem no quarto, e por analogia o computador, está executando uma manipulação puramente sintática (formal) de símbolos não interpretados. Ele tem as regras, mas não o significado. A compreensão, a semântica e a intencionalidade intrínseca são fenômenos que não podem emergir de uma manipulação puramente formal. Searle (2021) também refuta as objeções mais comuns, como a "objeção dos sistemas" (que afirma que o sistema como um todo entende) e a "objeção do robô" (que adiciona sensores e atuadores), argumentando que a adição de mais inputs sintáticos não resolve a falta fundamental de compreensão semântica no núcleo do processamento

A seguir as definições e diferenças entre IA Forte e IA Fraca e a posição de Searle:

Quadro 3 – Diferenças entre IA Fraca e IA Forte

Característica	IA Fraca (ou Cautelosa)	IA Forte
Papel do computador	O computador é uma ferramenta poderosa para o estudo da mente, permitindo simular processos e testar hipóteses de forma rigorosa.	O computador não é apenas uma ferramenta; um computador adequadamente programado é, literalmente, uma mente.

Status do programa	Os programas são instrumentos que nos capacitam a testar explicações psicológicas.	Os programas não são meros instrumentos, mas constituem as próprias explicações da cognição. São, em si, teorias psicológicas.
Afirmação sobre a mente	Não afirma que o computador possui estados mentais genuínos. Ele apenas simula capacidades cognitivas para fins de estudo.	Afirma que um computador adequadamente programado literalmente tem estados cognitivos, como compreensão, crenças e desejos.
Pressuposto fundamental	O computador é um modelo útil para a mente.	A mente está para o cérebro assim como o programa (software) está para o computador (hardware).
Posição de Searle	Searle não tem objeções a essa abordagem, considerando-a um instrumento valioso para a pesquisa.	Searle argumenta que essa tese é falsa, usando o Argumento do Quarto Chinês para demonstrar que a manipulação de símbolos (sintaxe) não é suficiente para a compreensão (semântica).

Fonte: elaborado pelo autor.

2.8 O contraponto funcionalista: a “atitude intencional” de Daniel Dennett

Uma alternativa radical à visão de Searle é a abordagem de Dennett. Para Dennett (1987), a intencionalidade não é uma propriedade biológica intrínseca e real de um sistema, mas sim uma estratégia preditiva que nós, como observadores, adotamos. Essa é a “atitude intencional” (*Intentional Stance*). Nós tratamos um sistema – seja uma pessoa, um animal, um termostato ou um computador de xadrez – *como se* ele fosse um agente racional com crenças e desejos, a fim de prever seu comportamento de forma eficiente.

Um sistema é um “verdadeiro crente” (*true believer*) se seu comportamento é previsto de forma confiável e robusta a partir da atitude intencional (Dennet, 1987). Para Dennett, não

existe um "fato mais profundo" sobre se o sistema *realmente* possui crenças em sua cabeça. A intencionalidade é um padrão objetivo no comportamento do sistema, mas um padrão que só é visível a partir dessa perspectiva interpretativa.

Dennett (1994) estende essa lógica à própria intencionalidade humana. Ele argumenta que nossa intencionalidade também é, em última análise, derivada. Utilizando a noção dos "genes egoístas" de Richard Dawkins, ele sugere que somos "máquinas de sobrevivência" projetadas pela "mãe natureza" (o processo de seleção natural). A intencionalidade que exibimos é uma característica funcional que foi selecionada por sua utilidade. Assim, a distinção fundamental de Searle entre intencionalidade "original" (a nossa) e "derivada" (a dos artefatos) é dissolvida. No final, toda intencionalidade é derivada de um processo de design, seja ele o de um engenheiro humano ou o da seleção natural.

2.9 Por que prevalecer com a teoria de Searle

Apesar da elegância da teoria de Dennett, a abordagem de Searle é adotada nesta dissertação por razões cruciais que se alinham diretamente com a questão central da pesquisa.

Primeiramente, a abordagem de Dennett, sendo uma estratégia de terceira pessoa focada na previsão de comportamento, não consegue acomodar adequadamente a realidade ontológica da consciência e da subjetividade (Searle, 2006). Para Searle (2006), a experiência de primeira pessoa – o "sentir-se como" de um estado mental – é um fato biológico irredutível, o fundamento da mente. A teoria de Dennett (1987), ao tratar a intencionalidade como uma atribuição externa, corre o risco de tornar a experiência subjetiva um epifenômeno ou uma "ilusão do usuário", algo que a teoria de Searle (2006) se recusa a fazer, insistindo na sua primazia causal e ontológica.

Em segundo lugar, para a questão central desta dissertação – a possibilidade de intencionalidade em sistemas não biológicos –, é necessário um critério de demarcação causal, não apenas preditivo. A "atitude" de Dennett é um critério de previsão comportamental. Se um robô se comportar de forma suficientemente complexa, poderíamos adotar a atitude intencional em relação a ele, e ele se tornaria um "verdadeiro crente". Isso, no entanto, não responde se o robô *realmente* tem mente da mesma forma que um humano ou algo que o equivalha. A teoria de Searle (2006), ao contrário, fornece um critério ontológico e causal: os "poderes causais

específicos do cérebro". A questão torna-se: o sistema artificial pode replicar a *causalidade* biológica que produz a intencionalidade, e não apenas seu *efeito* comportamental?

Finalmente, a manutenção da distinção entre intencionalidade intrínseca e derivada é essencial para a própria formulação da pergunta da dissertação. Se, como sugere Dennett (1987), toda intencionalidade é derivada (de genes ou de programadores), então, a diferença entre a intencionalidade de um humano e a de um robô sofisticado é apenas de grau e complexidade, não de tipo. A questão da emergência da mente em máquinas perde sua força. A estrutura de Searle (2006), que insiste na distinção, é o que permite que a pergunta "pode a intencionalidade *intrínseca* emergir em sistemas não biológicos?" seja formulada de maneira rigorosa e significativa.

O quadro 4 a seguir sintetiza as divergências fundamentais que justificam a escolha teórica deste trabalho.

Quadro 4 – Divergências entre John Searle e Daniel Dennet

Característica	John Searle (Naturalismo Biológico)	Daniel Dennett (Instrumentalismo/Funcionalismo)
Status da intencionalidade	Propriedade biológica intrínseca e real, causada pelo cérebro (Searle, 2006, p. 134; Searle, 2002).	Uma estratégia preditiva (atitude/*stance*). Um padrão objetivo, mas discernível apenas a partir de uma perspectiva interpretativa (Dennett, 1987, p. 27).
Relação com a consciência	Intrinsecamente ligada. A intencionalidade inconsciente só é mental porque é, em princípio, acessível à consciência (Searle, 1990, p. 585).	A consciência é um problema separado. A intencionalidade pode ser atribuída a sistemas não conscientes (termostatos, computadores) (Dennett, 1987, p. 33).
Origem da intencionalidade	Produto da evolução biológica e dos poderes causais específicos do	Derivada do processo de design da seleção natural ("mãe natureza").

Característica	John Searle (Naturalismo Biológico)	Daniel Dennett (Instrumentalismo/Funcionalismo)
	cérebro (Searle, 2006, p. 134).	Não há intencionalidade "original" (Dennett, 1994, p. 104).
Intencionalidade em IA	Impossível apenas com programas (sintaxe). Requereria a replicação dos poderes causais do cérebro (Searle, 2021, p. 1).	Possível. Se o comportamento de um computador for melhor previsto pela atitude intencional, ele é um "verdadeiro crente" (Dennett, 1987, p. 33).
Critério de demarcação	Ontológico e causal (poderes causais do cérebro) (Searle, 2006, p. 137).	Funcional e preditivo (sucesso da atitude intencional) (Dennett, 1987, p. 27).
Realidade subjetiva	Fato central e irredutível da mente (ontologia de primeira pessoa) (Searle, 2006, p. 14, 326).	Problemática; tende a ser vista como um produto da complexidade funcional ou uma "ilusão do usuário" (Dennett, 1987).

Fonte: elaborado pelo autor.

2.10. Conclusão do capítulo

Este capítulo procurou estabelecer um fundamento teórico rigoroso para a investigação da intencionalidade, culminando na adoção da teoria do Naturalismo Biológico de Searle (1990; 2002). A definição de intencionalidade que guiará esta dissertação é, portanto, a seguinte: a intencionalidade intrínseca é uma propriedade biológica de nível superior, causada por e realizada em processos neurofisiológicos específicos do cérebro, caracterizada por uma estrutura interna (conteúdo, modo, direção de ajuste), sustentada por uma rede de outros estados mentais e por um *background* de capacidades pré-representacionais, e indissociavelmente ligada à capacidade de consciência, conforme o Princípio da Conexão.

A justificativa para essa escolha teórica reside em sua superioridade sobre alternativas, como o funcionalismo de Dennett, para os propósitos específicos desta pesquisa. A teoria de Searle fornece os únicos critérios – a causalidade biológica específica e a irredutibilidade da

ontologia de primeira pessoa – que permitem uma investigação rigorosa sobre a possibilidade de a intencionalidade transcender suas origens biológicas. Sem esses critérios, a distinção entre um ser genuinamente intencional e uma simulação perfeita se dissolve, tornando a questão central da tese intratável.

Com essa definição de intencionalidade firmemente estabelecida, o caminho está preparado para as investigações subsequentes. O capítulo 3 aplicará esse arcabouço, para explorar as evidências arqueológicas de comportamento intencional na pré-história, analisando artefatos e padrões de comportamento em termos de suas causas neurobiológicas e funções adaptativas, sem recorrer a atribuições mentais injustificadas. Subsequentemente, o capítulo 4 confrontará diretamente a questão da intencionalidade artificial, investigando se os sistemas artificiais atuais ou futuros poderiam, em princípio, satisfazer os rigorosos critérios biológicos e causais de Searle para a intencionalidade intrínseca, ou se estão, por sua própria natureza, confinados ao domínio da sintaxe e, consequentemente, da intencionalidade derivada.

3. ORIGENS DA INTENCIONALIDADE: UMA INVESTIGAÇÃO ARQUEOLÓGICA

Após termos estabelecido, no capítulo anterior, a intencionalidade como um fenômeno irredutivelmente biológico, fundamentado na teoria naturalista de Searle (1983; 1992), este capítulo se volta para o registro material em busca de suas origens. Se a intencionalidade é um produto da evolução, seus traços devem estar impressos nos artefatos deixados por nossos ancestrais. A Arqueologia, portanto, não serve aqui como um mero catálogo de tecnologias passadas, mas como uma fonte primária de evidências, ou uma "paleofisiologia do comportamento" (Leroi-Gourhan, 1964), para rastrear a longa jornada evolutiva da mente. Esse campo de investigação, hoje consolidado como Arqueologia Cognitiva, busca precisamente inferir os processos de pensamento de sociedades passadas a partir de sua cultura material (Renfrew; Zubrow, 1994).

Este capítulo trata da causalidade, verificando se a intencionalidade é um produto da cognição e da cultura, ou é a intencionalidade que produz e molda a cognição e a cultura. Argumentaremos que uma relação de causalidade linear é insuficiente. Em seu lugar, propomos um modelo de causalidade recíproca e coevolutiva, alinhado às teorias de coevolução gene-

cultura (Boyd; Richerson, 2005). Assim, uma capacidade intencional incipiente permite a realização de ações que modificam o ambiente; esse ambiente modificado (contexto cultural) cria novas pressões seletivas e oportunidades de aprendizado que favorecem o desenvolvimento de estruturas cognitivas mais complexas; por sua vez, uma cognição mais complexa permite o surgimento de estados intencionais mais sofisticados, reiniciando o ciclo em um novo patamar.

Para desvendar essa espiral evolutiva, primeiro definiremos o conceito de cognição e sua inter-relação indissociável com a intencionalidade. Em seguida, analisaremos o registro arqueológico, desde as primeiras indústrias líticas, para identificar os marcadores de diferentes formas de intencionalidade – distinguindo, conforme a teoria de Searle (1983), a intenção prévia da intenção-na-ação. Por fim, validaremos o argumento central de que (i) a evolução biológica foi diretamente beneficiada pela destreza técnica, e (ii) que a dimensão social, através da intencionalidade compartilhada, foi um elemento-chave para a expansão da própria capacidade intencional, ao construir o *background* e a rede sobre os quais nossos estados mentais operam (Searle, 1992).

3.1. A arqueologia da mente: inferindo a intenção a partir da pedra

Para investigar a mente no passado, é imperativo primeiro delinear seus conceitos fundamentais. Entendemos por cognição o conjunto de processos mentais através dos quais os organismos adquirem, processam, armazenam e utilizam informações para compreender e interagir com o mundo. Tais processos incluem percepção, atenção, memória, raciocínio, resolução de problemas e planejamento (Ganascia, 1996). A cognição não é, portanto, uma entidade única, mas um sistema multifacetado de capacidades que permite a um ser vivo navegar seu ambiente de forma adaptativa. Como destaca Thomas Wynn (2002), a arqueologia pode acessar esses processos ao identificar atributos nos artefatos que exigem mecanismos cognitivos específicos, como a memória de trabalho ou o planejamento hierárquico.

A relação entre cognição e intencionalidade, sob a ótica de Searle (1983), é de interdependência fundamental. A intencionalidade não existe no vácuo, ela é uma propriedade de determinados estados cognitivos. Um estado mental, como uma crença, um desejo ou uma intenção, só pode ser “sobre” algo porque ele é, em si, um estado hospedado em um sistema cognitivo. Deste modo, a cognição é o substrato necessário para a existência da

intencionalidade. Contudo, a relação é bidirecional: a intencionalidade também dirige e organiza a cognição. Uma *intenção* de construir uma ferramenta, por exemplo, mobiliza e estrutura uma cascata de processos cognitivos: a atenção se foca na matéria-prima, a memória de trabalho retém o plano, o raciocínio avalia o próximo golpe e o sistema motor executa a ação. A arqueologia, ao focar na intencionalidade, busca precisamente reconstruir essa organização da ação no passado (Ribeiro, 2022).

Portanto, a intencionalidade é tanto uma característica *da* cognição quanto um princípio organizador *para* a cognição. Ao analisar o registro arqueológico, não buscamos um "fóssil da intenção" isolado, mas sim as evidências de sistemas cognitivos em funcionamento, cuja complexidade crescente nos informa sobre a evolução da própria capacidade de direcionar a mente para o mundo. A ferramenta de pedra se torna, assim, um testemunho da interação entre a capacidade de agir (intenção) e a capacidade de pensar (cognição).

3.2. As primeiras evidências: da percussão oportunista à ferramenta sistematizada

A transição crucial, que marca o início da linhagem tecnológica humana, é a "invenção da tecnologia", o momento em que se passa a modificar deliberadamente um objeto para um uso futuro, superando o uso oportunista de objetos não modificados (Beaune, 2004). A indústria Olduvaiense (~2.6 Ma), associada geralmente ao *Homo habilis*, representa o primeiro exemplo claro e sistemático dessa transição. As ferramentas são morfologicamente simples – seixos lascados (*choppers*) e lascas –, mas sua produção, quando analisada pela metodologia da Cadeia Operatória (Lemonnier, 1992; Moreno de Sousa, 2019), revela uma estrutura intencional que já pode ser dissecada com o instrumental de Searle.

A ação de golpear um seixo com um percutor para produzir uma lasca afiada é um exemplo claro de intenção-na-ação. Conforme Searle (1983), essa é a forma de intencionalidade presente na própria ação, em que a consciência está direcionada para a execução do movimento presente com o objetivo imediato de causar uma fratura. Contudo, a Cadeia Operatória Olduvaiense revela mais do que isso. A seleção de matérias-primas específicas (Toth, 1985), por vezes transportadas por distâncias consideráveis dos locais de seu descarte, implica a existência de uma intenção prévia. O hominínio formou, em sua mente, um plano anterior e

distinto da ação: “Vou buscar aquele tipo de rocha para, depois, produzir ferramentas”. Essa capacidade de formular um plano, retê-lo na memória e executá-lo posteriormente representa um patamar cognitivo superior à mera resposta a um estímulo imediato. No Olduvaiense, portanto, já observamos a interação entre o plano (intenção prévia) e sua execução passo a passo (as múltiplas intenções-na-ação), marcando o limiar da cognição projetiva.

3.2.1 Lomekwi 3 (3,3 milhões de anos): um prólogo debatido

As evidências mais antigas de modificação de pedra vêm do sítio de Lomekwi 3, no Quênia, datadas em 3,3 milhões de anos. As ferramentas do sítio de Lomekwi 3 são grandes, pesadas e foram produzidas por uma técnica de percussão bipolar (bater uma rocha sobre outra apoiada em uma bigorna). Os prováveis autores seriam hominínios como o *Australopithecus afarensis*. Embora a produção dessas peças seja indubitavelmente intencional, seu significado cognitivo é objeto de intenso debate. A tecnologia de Lomekwi parece ser um evento isolado, sem continuidade temporal ou geográfica, o que levou alguns pesquisadores a sugerir que poderia representar uma “falsa partida” tecnológica ou uma solução local que não se consolidou em uma tradição cultural transmitida (Wynn *et al*., 2018).

3.2.2 A indústria Olduvaiense (~2,6 Ma): a sistematização da intenção

Uma mudança significativa ocorre por volta de 2,6 milhões de anos com o surgimento da indústria Olduvaiense, associada principalmente ao *Homo habilis*. As ferramentas olduvaienses, embora morfologicamente simples (seixos lascados, ou *choppers*, e as lascas resultantes), demonstram uma clara sistematização. Há uma compreensão controlada do ângulo de percussão e da fratura conchoidal da rocha para produzir gumes afiados de forma recorrente. A produção não é aleatória; o objetivo não é apenas quebrar a pedra, mas produzir lascas cortantes – o verdadeiro produto intencional – a partir de um bloco de rocha (o núcleo).

Aqui, a Cadeia Operatória revela um planejamento mais estruturado:

1. *Seleção e transporte de matéria-prima*: os hominínios selecionavam rochas com boas propriedades de fratura e, em muitos casos, as transportavam por distâncias consideráveis.

2. *Produção controlada*: a aplicação da percussão direta para remover uma série de lascas de um núcleo demonstra a compreensão da relação causa-efeito e a retenção de um objetivo (produzir gumes).
3. *Uso das lascas*: as lascas, e não apenas os núcleos, eram os principais instrumentos, utilizados para cortar carcaças e processar plantas.

A indústria Olduvaiense representa, portanto, a primeira tradição tecnológica duradoura e geograficamente expandida. Ela evidencia uma intencionalidade que não é apenas momentânea, mas que envolve antecipação (transportar a rocha para um local de uso futuro) e conhecimento técnico transmitido.

Contudo, descobertas recentes forçam uma revisão profunda desse cronograma e, mais importante, dos agentes envolvidos. Pesquisas no sítio de Nyayanga, no Quênia, recuaram a data de início da tecnologia Olduvaiense para um período entre 3.032 e 2.581 Ma (Plummer *et al*., 2023). Essa nova datação não é apenas um ajuste cronológico de quase meio milhão de anos, ela dissocia o surgimento da tecnologia lítica do período de significativa expansão cerebral no gênero *Homo* (Plummer *et al*., 2023), o que sugere que as bases cognitivas para a intencionalidade lítica – a capacidade de formular um plano e executá-lo para modificar o mundo material para um fim futuro – precedem o desenvolvimento de cérebros maiores que caracterizariam o *Homo erectus* posterior.

Ainda mais disruptiva é a associação contextual dessas ferramentas antigas. Em Nyayanga, artefatos Olduvaienses foram encontrados em associação direta com fósseis de *Paranthropus*, um gênero de hominínios que evoluiu em paralelo ao gênero *Homo* e que tradicionalmente não era considerado um fabricante de ferramentas (Plummer *et al*., 2023). A presença de um molar de *Paranthropus* em um contexto de abate de megafauna (hipopótamos), juntamente com as ferramentas usadas para processá-los, quebra o paradigma Homo-cêntrico da tecnologia lítica (Plummer *et al*., 2023). A intencionalidade, manifestada através da fabricação e do uso de ferramentas, pode não ter sido uma característica exclusiva da nossa linhagem direta.

Essa revelação abre duas possibilidades interpretativas com profundas implicações. A primeira é a de que a capacidade para a intencionalidade lítica é um traço ancestral, presente no ancestral comum do *Homo* e do *Paranthropus*, que foi subsequentemente expresso em ambas as linhagens. A segunda é a de que se trata de um caso de evolução convergente, em que

pressões seletivas semelhantes (por exemplo, a necessidade de acessar novos recursos alimentares em ambientes de savana) levaram ao desenvolvimento independente de soluções tecnológicas e cognitivas análogas em linhagens distintas.

Para a questão central desta dissertação, esse *insight* é fundamental. Se a intencionalidade biológica já demonstrou, em sua própria história evolutiva, o potencial para múltiplas emergências ou para uma herança mais ampla do que se supunha, o argumento de sua irredutibilidade a um único substrato (o cérebro do *Homo sapiens*) é enfraquecido. Isso sugere que a intencionalidade pode ser vista como uma solução adaptativa que emerge quando certas condições de complexidade neurológica, necessidade ecológica e oportunidade material são atendidas.

3.2.3. Decodificando a intenção na pedra

A atribuição de intencionalidade a agentes do passado, um dos pilares da Arqueologia Cognitiva, exige um rigor metodológico que transcende a simples intuição ou a interpretação anedótica. Para inferir um estado mental a partir de um objeto material, é necessário um arcabouço analítico que possa sistematicamente desmembrar o processo de criação do artefato em uma série de decisões e ações observáveis. É a partir da estrutura dessas decisões que podemos começar a reconstruir os planos e as intenções que as guiaram (referência?).

Este trabalho adota o conceito de Cadeia Operatória como sua principal ferramenta analítica para a investigação das indústrias líticas.

Originado nos trabalhos do etnólogo e arqueólogo francês André Leroi-Gourhan (1964) e posteriormente sistematizado por seus discípulos, notadamente Pierre Lemonnier (1992), o método da Cadeia Operatória reconstrói a "história de vida" de um artefato, desde a concepção mental até o seu descarte final. Conforme detalhado por Moreno de Sousa (2019), esse processo pode ser segmentado em estágios observáveis, cada um representando um nó de decisão que reflete o plano do artífice:

1. *Aquisição da matéria-prima*: a seleção de uma rocha específica envolve conhecimento sobre geologia, sobre as propriedades de fratura dos materiais e sobre a paisagem. A escolha revela uma intenção baseada na antecipação das necessidades do processo de lascamento.
2. *Etapas de produção*: esse estágio é subdividido em:

- *Formatação do núcleo*: modificação inicial do bloco para preparar uma superfície de percussão.
- *Debitagem*: a produção de lascas a partir do núcleo, que podem ser o produto final ou suportes para outras ferramentas.
- *Façonagem*: o processo de dar forma a um suporte (uma lasca ou o próprio núcleo) através de lascamentos sucessivos para criar um instrumento como um biface.
- *Retoque*: modificações finas nas bordas da ferramenta para afiá-la ou dar-lhe uma forma específica.

3. *Utilização*: o uso da ferramenta, que pode ser inferido por meio de análises de microdesgaste (traceologia).
4. *Descarte ou abandono*: o local e a forma como a ferramenta é descartada podem fornecer informações sobre a organização do espaço e as atividades realizadas.

Cada uma dessas etapas representa um conjunto de gestos técnicos e escolhas que refletem um plano mental. A seleção de um tipo específico de rocha, a técnica de percussão utilizada e a morfologia final da peça são "nós de decisão" que revelam o nível de planejamento, conhecimento técnico, tradição cultural e antecipação do hominínio. A Cadeia Operatória, portanto, como destacado por Moreno de Sousa (2019), permite-nos mover da descrição estática do artefato para a reconstrução dinâmica do comportamento, oferecendo uma janela para a mente do artífice e para as intenções que guiaram suas ações.

É importante notar que a opção pela Cadeia Operatória, com sua ênfase antropológica nas intenções, nos gestos e no conhecimento técnico, não invalida outras abordagens para o estudo de indústrias líticas, que oferecem contribuições valiosas para diferentes questões de pesquisa. Entre as principais, destacam-se:

- *Reduction Sequence* (Sequência de Redução): de tradição norte-americana e ligada à Arqueologia Processual, esse método foca de maneira mais empírica nas fases observáveis da redução de um núcleo lítico. Seu objetivo é quantificar as estratégias de lascamento e a eficiência na produção de artefatos, sem necessariamente recorrer a uma teoria externa sobre as intenções do artífice.
- *Mínimo Nódulo Analítico (MANA):* desenvolvido no Brasil por André Prous, baseia-se na identificação macroscópica da matéria-prima para remontar fragmentos de um

mesmo nódulo original. É particularmente poderoso para fazer inferências sobre a gestão de recursos, a mobilidade dos grupos caçadores-coletores e a organização espacial das atividades dentro de um sítio.

- *Gihō* (Técnica): com origem na arqueologia japonesa, esse método descritivo e culturalista é focado na reconstrução detalhada das sequências de produção de tipos específicos de artefatos, como as microlâminas. É extremamente útil para a definição de tipos, a identificação de tradições culturais e o estabelecimento de cronologias relativas.

O Quadro 5 abaixo resume as características e os focos de cada abordagem:

Quadro 5 – Comparativo de métodos de análise lítica

Método	Origem	Foco principal	Vantagens para o estudo da intencionalidade
Cadeia Operatória	França (Leroi-Gourhan)	Processo completo: intenções, gestos, conhecimento e contexto social.	Permite reconstruir o plano mental e as decisões do artífice em cada etapa.
Reduction Sequence	EUA (Processualismo)	Fases empíricas da redução do núcleo.	Foco objetivo nas transformações materiais; útil para quantificar estratégias.
MANA	Brasil (A. Prous)	Gestão da matéria-prima a partir da análise macroscópica.	Inferências sobre planejamento de mobilidade e economia de recursos.

Gihō	Japão	Sequências de produção de tipos específicos (e.g., microlâminas).	Forte poder descritivo para definir tradições culturais e cronologias.

Fonte: elaborado pelo autor com base nas referências indicadas anteriormente.

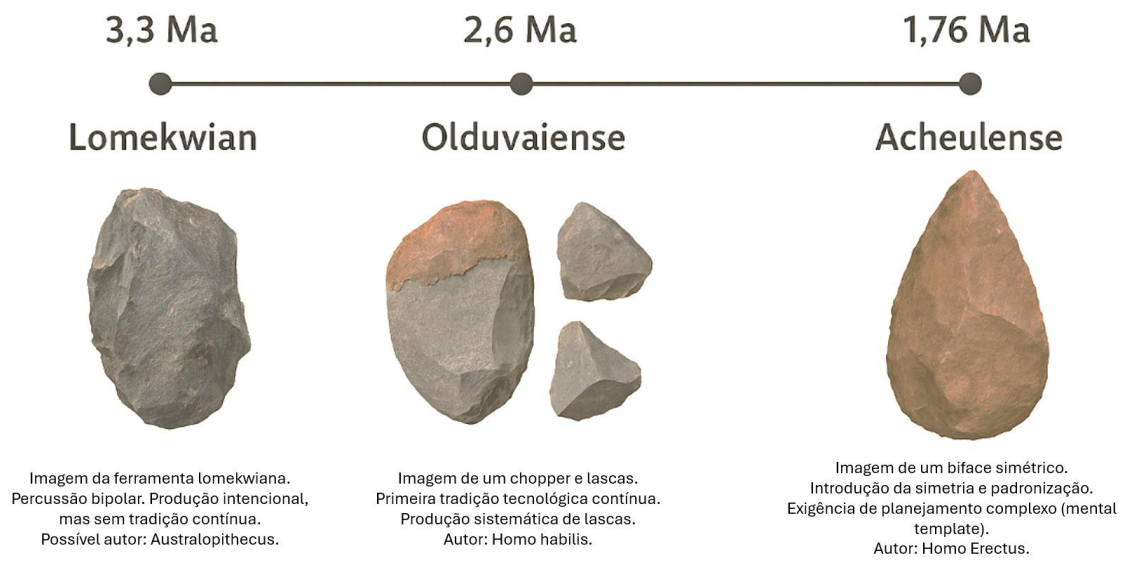
Apesar da validade de cada um desses métodos para seus respectivos objetivos, a escolha pela Cadeia Operatória nesta dissertação é deliberada. Para a questão central deste capítulo – a investigação da origem e da evolução da intencionalidade –, a perspectiva da Cadeia Operatória se mostra a mais adequada. Seu foco explícito na reconstrução das intenções, dos planos mentais e do conhecimento técnico que orientam a ação (Moreno de Sousa, 2019) a torna a ferramenta metodológica mais potente para conectar o artefato material (o objeto de estudo da arqueologia) ao estado mental (o objeto de estudo da filosofia da mente de Searle).

3.3. A consolidação da intencionalidade: simetria e padronização no Acheulense

Com a indústria Acheulense (~1.76 Ma), com origem associada ao *Homo erectus*, a evidência para uma intencionalidade complexa se torna robusta.

A produção de um biface simétrico e padronizado não pode ser explicada como um resultado acidental de ações simples. Ela exige a existência de uma intenção prévia muito bem definida e detalhada. Isso exigiu, como argumenta Thomas Wynn (1995; 2002), a manutenção de um "molde mental" (*mental template*) da forma final desejada na memória de trabalho.

Figura 2 – Linha do tempo da evolução tecnológica paleolítica



Fonte: elaborado pelo autor.

Esse plano mental prévio precisa ser mantido ativo ao longo de uma cadeia operatória longa e complexa, guiando centenas de ações individuais. Cada golpe é uma intenção-na-ação que não busca apenas lascar a pedra, mas aproximar a forma bruta da forma idealizada no plano. O artífice precisa avaliar a peça, rotacioná-la e antecipar o resultado de seus gestos em um espaço tridimensional. O biface Acheulense é, portanto, a materialização de uma intencionalidade hierárquica, em que uma intenção prévia geral governa uma multitude de intenções-na-ação subordinadas. É nesse ponto que a filosofia da técnica de Simondon (1980) se torna particularmente relevante: o artífice, ao impor uma forma à matéria através de um processo disciplinado, não apenas "individua" o objeto técnico, tornando-o coerente em sua estrutura, mas também passa por um processo de "subjetivação", em que sua própria cognição e subjetividade são moldadas e complexificadas pela relação com a técnica (Simondon, 1980).

Diferentemente da produção de lascas olduvaienses, na qual o objetivo principal é o produto imediato, a confecção de um biface requer a imposição de uma forma tridimensional preconcebida sobre a matéria-prima. Thomas Wynn (1995) argumenta que a simetria bilateral observada em muitos bifaces acheulenses não possui uma justificativa puramente funcional. A simetria seria, então, um atributo imposto ao objeto por razões que transcendem a utilidade imediata, refletindo uma capacidade cognitiva de trabalhar com conceitos espaciais complexos e uma manifestação de uma convenção estilística (Shipton *et al.*, 2018). A sua forma

padronizada, encontrada em vastas áreas geográficas e por mais de um milhão de anos, indica a adesão a uma norma compartilhada, um *Bauplan* mental que guiava a produção (Corbey *et al*., 2016). Alguns exemplares são tão grandes e finamente trabalhados que seu uso prático é questionável, levando à especulação de que eram valorizados por sua aparência e poderiam ter funcionado como um sinal social ou de aptidão, uma demonstração de habilidade do artífice (Shipton *et al*., 2018).

A existência de uma convenção estilística duradoura é a primeira evidência arqueológica robusta de uma norma social (Corbey *et al*., 2016). A capacidade de um grupo de hominínios de compartilhar, transmitir e aderir a uma regra abstrata ("o objeto deve ser simétrico") é o alicerce cognitivo para a intencionalidade coletiva (Searle, 1995). Esse conceito, central na filosofia de Searle (1995) descreve a capacidade de indivíduos agirem juntos com uma consciência compartilhada de seus objetivos e papéis. A produção de um biface simétrico não é, portanto, apenas um ato técnico individual, é um ato de conformidade social. O artífice está alinhando seu comportamento a uma regra do grupo, participando de uma realidade socialmente construída. O biface simétrico pode ser visto como um "fato institucional" proto-histórico, no qual a intencionalidade coletiva impõe uma função (nesse caso, uma forma valorizada) a um objeto (Searle, 1995). Isso marca um ponto de inflexão na evolução da mente, em que a intencionalidade transcende o indivíduo e se torna uma propriedade do grupo cultural.

A Cadeia Operatória de um biface é substancialmente mais longa e complexa, pois envolve centenas de golpes controlados e requer a rotação constante da peça e a preparação de plataformas de percussão, exigindo que o artífice mantenha em sua mente uma "imagem mental" (*mental template*) do produto final ao longo de todo o processo. Esse *mental template* é uma manifestação clara do polo mental no modelo de Ganascia (1996): uma representação simbólica e procedural que guia a ação motora complexa. No contexto dessa dissertação, o exemplo mais claro do polo mental é o "molde mental" (*mental template*) que um hominínio precisava ter para fabricar um biface Acheulense. Esse molde não é o biface físico (que pertence ao polo cultura) nem um conjunto de neurônios específicos (do polo cérebro), mas sim um plano representacional e procedural – uma estrutura de informação – que guia a ação intencional.

Esse processo evidencia uma forma de intencionalidade prospectiva, conceito que encontra profunda ressonância nas teorias de Simondon (1980). A intencionalidade prospectiva é um conceito que descreve uma forma de intenção, materializada no design de um objeto, que

antecipa e projeta o seu potencial de desenvolvimento e uso futuro. Diferentemente de uma intenção voltada para um fim imediato, ela se inscreve na estrutura do objeto como uma solução técnica aberta, com uma trajetória de existência futura já projetada em sua concepção. Alinhada à filosofia de Simondon (1980), ela é exemplificada pelo biface Acheulense, cuja forma versátil e durável sugere um plano que transcende a necessidade de um simples gume afiado. A intencionalidade prospectiva não contradiz a teoria de John Searle – pelo contrário, ela pode ser entendida como uma manifestação particularmente complexa e evolutivamente significativa das estruturas que Searle descreve, a intenção prévia é o plano mental formado antes da ação (ex: "vou fazer um biface"). Lembrando ainda que John Searle argumenta que a intencionalidade opera dentro de uma rede (*network*) de outras crenças e desejos. Uma intencionalidade prospectiva exige e ao mesmo tempo fomenta uma rede extremamente rica. Para que um hominínio formasse a intenção prospectiva de "criar um biface", ele necessitaria de uma vasta rede de crenças e desejos. Adicionalmente, remetendo ainda a Searle (1983), a execução de um plano prospectivo complexo, como o de fazer um biface, só é possível mediante um robusto *background* de capacidades: a habilidade motora fina, a postura corporal automática, o "sentir" da pedra, a coordenação olho-mão.

A prática contínua de produzir objetos com intencionalidade prospectiva é um dos principais mecanismos que constroem e enraízam o *background* técnico. Através da aprendizagem e da repetição (ou seja, da transmissão cultural), essas habilidades se tornam "segunda natureza", deixando de ser um esforço consciente e passando a integrar o *background* do indivíduo e do grupo. (Searle, 1983)

A intencionalidade prospectiva pode ser vista como a manifestação, no domínio técnico e evolutivo, de uma estrutura intencional Searliana que atingiu um alto grau de complexidade. Ela representa uma intenção prévia que se apoia em uma rede de crenças complexas e só é realizável através de um *background* de habilidades corporais sofisticadas, demonstrando como a abstração filosófica de Searle (1983) encontra uma correspondência concreta na evolução da cognição e da cultura material humana.

Para além da busca pela "concretude" do objeto, a produção do biface pode ser entendida através do processo de individuação técnica (Simondon, 1980). A individuação não é meramente a fabricação de um item, mas o processo pelo qual um objeto técnico se torna ele mesmo, alcançando uma coerência interna em que forma, material e função estão tão integrados

que o objeto passa a funcionar como um "indivíduo" autocondicionado. O biface, ao superar a condição de "rocha quebrada" e adquirir uma norma interna de simetria e equilíbrio, está se individuando. Contudo, o aspecto mais profundo da teoria de Simondon (1980) reside na reciprocidade desse processo: a subjetivação. Ao individuar o objeto técnico, o sujeito humano também se individua. A disciplina, o planejamento, a antecipação de gestos, a memória procedural e a atenção sustentada exigidos para criar um biface não são apenas habilidades aplicadas, mas processos que constituem e complexificam o próprio sujeito. O hominínio não apenas faz a ferramenta, ele se faz na relação com ela. A intencionalidade projetada no objeto reverbera, transformando a estrutura cognitiva e a subjetividade de seu criador.

3.4. A dimensão social da mente: intencionalidade individual vs. compartilhada

A notável padronização dos bifaces Acheulenses e sua persistência por mais de um milhão de anos em três continentes seriam inexplicáveis se a habilidade para os produzir fosse redescoberta a cada geração. A existência de uma "tradição" Acheulense, com normas estilísticas e técnicas que transcendem o indivíduo, aponta inequivocamente para uma dimensão social da cognição e para a existência de mecanismos eficientes de transmissão de conhecimento. Para analisar essa dimensão, é crucial recorrer ao quadro evolutivo da cognição social proposto por Michael Tomasello (1999), que estabelece uma distinção fundamental entre a intencionalidade puramente individual e as formas exclusivamente humanas de intencionalidade compartilhada.

Para compreender o salto evolutivo, devemos primeiro definir a linha de base. A intencionalidade individual é a capacidade, presente em muitos animais e certamente nos grandes primatas, de um organismo ter seus próprios objetivos e de formular planos para alcançá-los. Um chimpanzé que usa um graveto para capturar cupins age com intencionalidade individual, ele pode até aprender a técnica observando outro (um processo conhecido como emulação, em que se copia o resultado, mas não necessariamente o processo ou a intenção do modelo), mas seu objetivo e seu plano permanecem estritamente seus (Tomasello, 2014). Essa forma de cognição permite uma adaptação flexível ao ambiente, mas não cria a base para a cultura cumulativa que vemos na linhagem humana.

O grande divisor de águas na evolução da cognição na linhagem humana (gênero *homo*) é o surgimento da intencionalidade compartilhada, um termo guarda-chuva para a capacidade de formar "nós-intenções" (*we-intentions*). Tomasello (2014) argumenta que essa capacidade não surgiu de uma só vez, mas evoluiu em duas grandes etapas. A primeira, e mais relevante para o contexto do *Homo erectus* e da indústria Acheulense, é a intencionalidade conjunta (*Joint Intentionality*). Essa é a capacidade de dois ou mais indivíduos, em uma interação direta e em tempo real, formarem uma meta comum e coordenarem suas ações e papéis para atingi-la. A intencionalidade conjunta não é apenas a soma de duas intencionalidades individuais, ela requer um novo maquinário cognitivo-social, cujos componentes são (1) a formação de uma meta conjunta, na qual ambos os participantes estão comprometidos com o sucesso mútuo; (2) a atenção conjunta, na qual ambos sabem que estão focados no mesmo objeto ou objetivo; e (3) a capacidade de compreender e inverter papéis, em que um participante entende a perspectiva e a função do parceiro e pode, se necessário, assumir seu papel para garantir o sucesso da empreitada (Tomasello, 2008).

A transmissão do conhecimento para fabricar um biface Acheulense transcende a mera imitação e entra plenamente no domínio da intencionalidade conjunta. A aprendizagem não poderia ocorrer apenas pela observação do resultado final (emulação), dada a complexidade do processo. Ela exigiria um cenário de instrução social, no qual um mestre e um aprendiz compartilham a meta conjunta de replicar a forma do artefato. O aprendiz precisaria compreender a intenção por trás dos gestos do mestre, e o mestre, por sua vez, precisaria monitorar o estado de atenção e compreensão do aprendiz, ajustando sua demonstração. Esse cenário de ensino e aprendizagem é uma forma paradigmática de intencionalidade conjunta (Tomasello, 2014).

É precisamente através dessas interações de intencionalidade conjunta que podemos compreender, em termos Searlianos, a construção do *background* e da rede que tornam a intencionalidade técnica complexa possível. Conforme Searle (1983; 1992), nossos estados intencionais só funcionam porque estão ancorados em um *background* de saber-fazer, habilidades corporais e posturas pré-intencionais. A prática guiada e a aprendizagem social na fabricação de ferramentas constroem e solidificam esse *background* técnico no indivíduo e no grupo. Da mesma forma, a intenção de "fazer um biface" só faz sentido dentro de uma rede de crenças (sobre a qualidade da pedra, sobre a forma correta) e desejos (de ter uma ferramenta

eficaz, de pertencer ao grupo), e essa rede é alimentada e mantida pela cultura compartilhada. A intencionalidade conjunta, portanto, é o mecanismo psicossocial que permite a criação e a transmissão do contexto cultural (*background* e rede) que, por sua vez, capacita e amplifica a intencionalidade individual de cada novo artífice.

Posteriormente na linhagem humana, já com os humanos anatomicamente modernos, a intencionalidade conjunta evoluiria para a intencionalidade coletiva. Essa forma mais abstrata não depende mais da interação diádica aqui-e-agora, mas da existência de um "terreno cultural comum" de normas, convenções e instituições. A intencionalidade coletiva é a capacidade de alinhar-se às regras e às expectativas de um grupo cultural anônimo, adotando uma perspectiva "agente-neutra". Ela explica a emergência de tradições simbólicas, rituais e linguagens convencionais. Enquanto a intencionalidade conjunta explica como um aprendiz aprende com um mestre a fazer um biface, a intencionalidade coletiva explica por que as "Vênus" do Gravetiense, espalhadas por milhares de quilômetros, seguem a mesma convenção estilística (Tomasello, 2014).

Para clarificar essas distinções fundamentais, o Quadro 6 a seguir resume as características de cada tipo de intencionalidade.

Quatro 6 – Características das intencionalidades individual, conjunta e coletiva

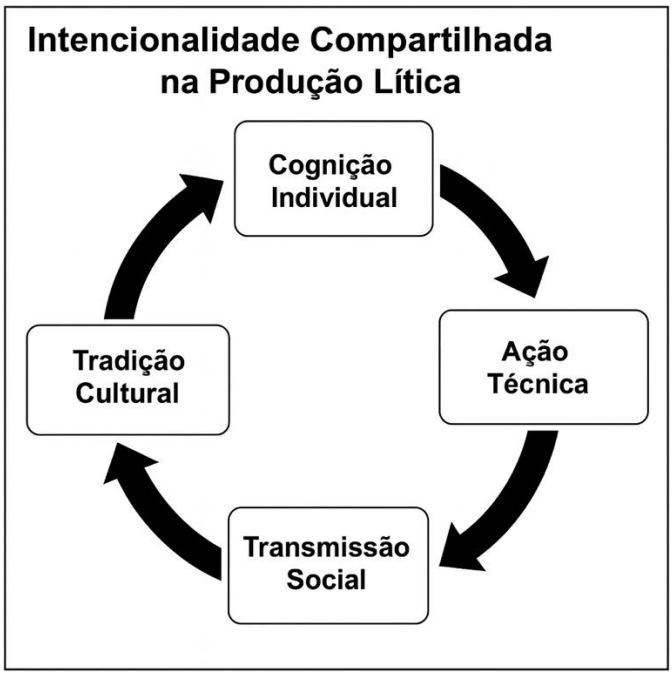
Característica	Intencionalidade individual	Intencionalidade conjunta	Intencionalidade coletiva
Agente	Eu	Eu e você (nós-diádico)	Nós (o grupo cultural)
Meta	Pessoal	Meta compartilhada em tempo real	Meta convencionalizada / Norma social
Perspectiva	Egocêntrica	Segunda pessoa (entender seu papel)	Agente-neutra / "objetiva" (entender as regras do grupo)
Contexto	Situação imediata	Interação direta aqui-e-agora	Terreno cultural comum (convenções)

Hominínio associado	Ancestral comum / Grandes primatas	Homo erectus (hipótese)	Humanos anatomicamente modernos
Exemplo arqueológico	Uso de uma pedra não modificada	Ensino/aprendizagem da técnica do biface	Produção de adornos e arte seguindo uma tradição regional

Fonte: elaborado pelo autor.

A transmissão do conhecimento para fabricar um biface transcende a simples imitação. Exige que o aprendiz compreenda a meta do mestre, não apenas seus movimentos. Ele precisa internalizar o plano, a intenção por trás de cada gesto técnico, em um processo que se assemelha a uma atividade colaborativa com o objetivo compartilhado de replicar a forma. Essa capacidade de formar uma "intenção-nós" ("nós vamos fazer um biface assim") é a fundação da transmissão cultural de alta fidelidade, como proposto por teóricos como Boyd e Richerson e ilustrada na Figura 3. A conformidade com a norma técnica não é apenas uma cópia, mas uma participação em uma tradição, um ato de alinhamento com a intencionalidade coletiva do grupo. A tecnologia lítica, portanto, deixa de ser apenas um produto da cognição individual para se tornar um fenômeno social, um repositório de conhecimento coletivo que conecta gerações (Moreno de Sousa, 2019).

Figura 3 – Diagrama da intencionalidade compartilhada na produção lítica



Fonte: elaborado pelo autor.

- *Cognição individual*: formação da intenção prévia (*Mental Template* do biface)
- *Ação técnica*: execução da Cadeia Operatória
- Transmissão social (intencionalidade compartilhada – Tomasello): aprendizagem através da compreensão da intenção do mestre (ensino/imitação)
- Tradição cultural (Boyd; Richerson, 2005): padronização e estabilidade da forma do artefato através de mecanismos de transmissão cultural**.**

Em benefício da profundidade desta dissertação, é crucial adicionar uma camada de nuance à discussão. A necessidade de transmissão cultural de alta fidelidade não deve ser generalizada para todas as tecnologias pré-históricas. Um estudo experimental recente de Snyder e *et al*. (2022) oferece um contraponto informativo fundamental. A pesquisa demonstrou que as técnicas de lascamento mais simples, características da indústria Olduvaiense, não necessitam de transmissão cultural de *know-how*. Participantes experimentalmente ingênuos, sem qualquer instrução, foram capazes de reinventar espontaneamente as técnicas de lascamento Olduvaiense para resolver um problema que exigia uma ferramenta de corte. As técnicas, portanto, "poderiam ter sido derivadas individualmente" (Snyder *et al*., 2022).

Esse achado não invalida a necessidade de aprendizagem social para tecnologias mais complexas como a Acheulense, mas refina nossa compreensão. Ele sugere que a presença de ferramentas simples no registro arqueológico não é, por si só, uma prova de ensino ou de transmissão cultural complexa. A combinação dos dados sobre a complexidade Acheulense e a simplicidade Olduvaiense aponta para uma coevolução dos mecanismos de transmissão cultural com a própria tecnologia. Para a tecnologia Olduvaiense, mecanismos de baixa fidelidade, como a emulação ou o aprimoramento de estímulo, podem ter sido suficientes para sua disseminação (Lycett, 2015). Para a tecnologia Acheulense, um salto para um protocolo de alta fidelidade – o ensino, fundamentado na intencionalidade conjunta – tornou-se necessário (Morgan *et al*., 2015).

Portanto, a transição Olduvaiense-Acheulense não foi apenas uma mudança tecnológica, mas também uma mudança no modo de transmissão social, refletindo uma transformação fundamental na estrutura da intencionalidade, da capacidade predominantemente individual para a emergência da intencionalidade conjunta. Posteriormente na linhagem humana, já com os humanos anatomicamente modernos, a intencionalidade conjunta evoluiria para a intencionalidade coletiva. Essa forma mais abstrata não depende mais da interação diádica em tempo real, mas da internalização de regras, normas e convenções sociais que governam o comportamento do grupo como um todo, como discutido na seção anterior sobre a simetria dos bifaces (Searle, 1995).

3.5. Da função à origem do significado: uma perspectiva paleoantropológica

A complexidade cognitiva e intencional evidenciada pela indústria Acheulense levanta uma questão fundamental: essa sofisticação técnica equivale à emergência do significado, no sentido simbólico do termo? A perspectiva paleoantropológica de Walter Neves e outros autores oferece uma resposta nuançada e cautelosa. Neves propõe uma distinção crucial entre cognição complexa e cognição simbólica (Neves *et al.*, 2020).

A produção de um biface é, como já discutido acima, um feito de cognição complexa (Wynn, 1995; Stout *et al*., 2008). Exige planejamento hierárquico, memória de trabalho, controle motor fino e uma sensibilidade estética para a simetria. Contudo, a forma do biface,

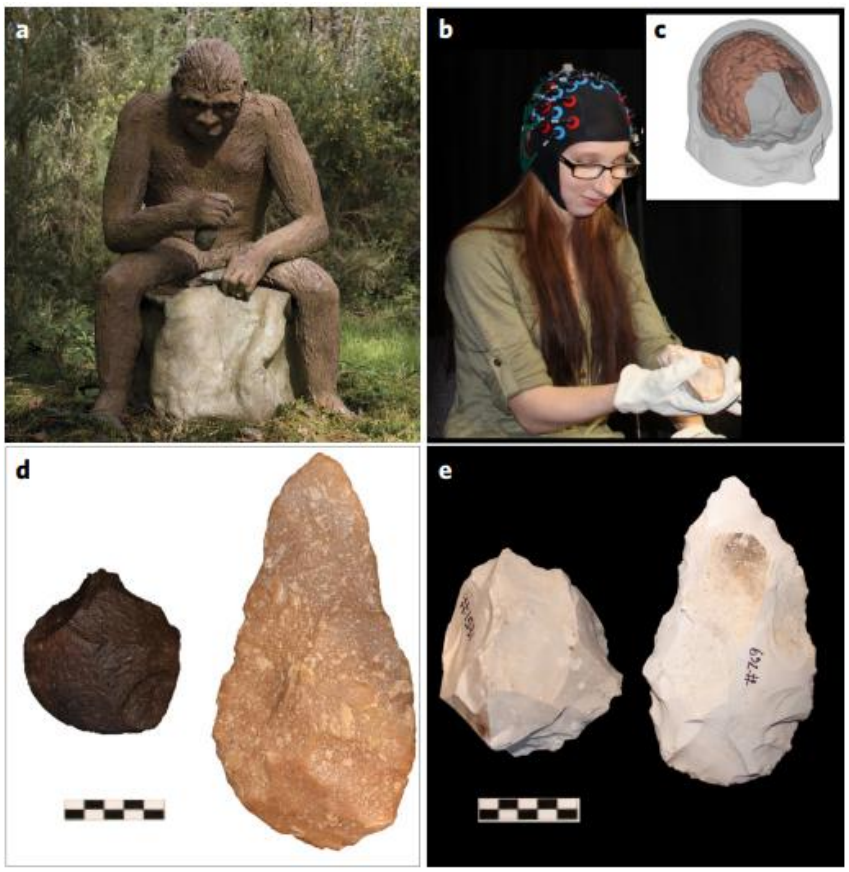
por mais padronizada que seja, permanece indexicalmente ligada à sua função e às propriedades do material. Sua "significação" é primariamente pragmática. A cognição simbólica, por outro lado, é definida pela capacidade de criar e manipular símbolos arbitrários, em que a relação entre o significante (a palavra "pedra") e o significado (o conceito do objeto pedra) é puramente convencional e estabelecida socialmente.

Segundo essa abordagem, embora *hominínios* como o *Homo erectus* e os *Neandertais* possuíssem uma mente técnica e socialmente complexa, a explosão da cognição simbólica – manifesta em arte rupestre, adornos pessoais e, presume-se, linguagem plenamente sintática – seria um fenômeno mais recente e uma característica distintiva dos humanos anatomicamente modernos. As capacidades cognitivas desenvolvidas ao longo de milhões de anos de fabricação de ferramentas, no entanto, não devem ser subestimadas. Elas foram os precursores necessários para o advento do significado. O cérebro capaz de gerenciar a sintaxe de ações para criar um biface estava estabelecendo as fundações neurais e cognitivas sobre as quais, mais tarde, a sintaxe da linguagem e o pensamento simbólico poderiam florescer. (Neves, 2020)

3.6. A mente no cérebro: evidências da neuroarqueologia e a coevolução da tecnologia e da linguagem

A argumentação de que a intencionalidade é um fenômeno biológico, enraizado na estrutura cerebral (Searle, 1983), encontra um campo de testes empíricos na Neuroarqueologia. Essa disciplina, pioneira em grande parte pelo trabalho de Dietrich Stout e seus colaboradores (Stout *et al.,* 2008), busca correlacionar a produção de artefatos com a atividade neural, oferecendo uma ponte entre o registro arqueológico e o polo do cérebro no modelo de Ganascia (1996). Utilizando técnicas como a Tomografia por Emissão de Pósitrons (PET) e a Imagem por Ressonância Magnética funcional (FMRI) representadas na Figura 4, esses estudos permitem observar quais redes cerebrais são recrutadas durante a fabricação de ferramentas de pedra, revelando as demandas cognitivas de cada tecnologia (Stout *et al*., 2008).

Figura 4 – Configuração experimental do processo de redução lítica



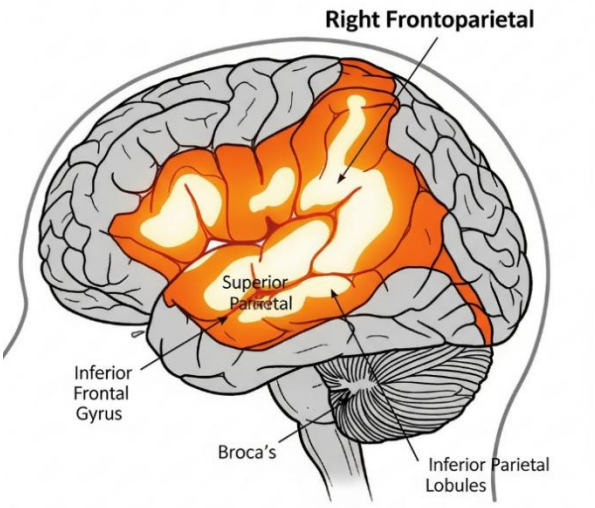
Fonte: The functional brain networks that underlie Early Stone Age tool manufacture (Putt, 2017). A-C – Os processos de redução lítica realizados por Homo arcaico (a) foram replicados por 31 participantes humanos modernos enquanto utilizávamos espectroscopia funcional no infravermelho próximo (b) para registrar a atividade cerebral regional em partes dos córtices frontal, parietal e temporal do cérebro (c). d,e, Ferramentas dos tipos Oldowan (d,e, à esquerda) e Acheuliana (d,e, à direita), oriundas do registro arqueológico (d), foram reproduzidas pelos participantes do estudo (e).

Os experimentos de Stout *et al.* (2008), comparando a fabricação de ferramentas Olduvaienses e Acheulenses por indivíduos novatos e especialistas, forneceram *insights* detalhados sobre a evolução da cognição humana.

Oldowan vs. Acheulense: a fabricação de ferramentas Olduvaienses por novatos ativa principalmente circuitos parietofrontais dorsais, que são evolutivamente antigos e homólogos aos sistemas de preensão visualmente guiada em primatas não humanos (Stout *et al.,* 2008). Notavelmente, não há uma ativação significativa do córtex pré-frontal dorsolateral, associado ao planejamento estratégico e à memória de trabalho de alto nível (Stout *et al*., 2008). Isso sugere que as demandas cognitivas da tecnologia Olduvaiense inicial estavam mais focadas na coordenação visuomotora imediata do que em um planejamento complexo e de longo prazo.

Em contraste, a fabricação de ferramentas Acheulenses por especialistas recruta uma rede neural muito mais ampla e complexa. A principal diferença é a ativação robusta do sistema frontoparietal ventral (Stout *et al.,* 2008). Crucialmente, isso inclui o giro pré-frontal inferior direito (BA 45), um homólogo da área de Broca no hemisfério esquerdo, que é fundamental para a linguagem (Stout *et al.,* 2008) representada na Figura 5. A ativação dessa região, associada ao planejamento hierárquico, ao controle cognitivo e à organização de sequências de ação complexas, indica que a tecnologia Acheulense impôs demandas cognitivas de uma ordem superior, alinhadas com a necessidade de seguir um plano mental pré-concebido e de executar centenas de gestos sequenciados com precisão (Stout *et al.,* 2008).

Figura 5 – Representação da rede frontoparietal direita ativada durante a fabricação de ferramentas Acheulenses



Fonte: elaborado pelo autor com base nos achados de Stout e Chaminade (2012).

Ilustração esquemática do hemisfério direito do cérebro humano. As áreas destacadas em laranja representam a rede neural cuja ativação mais robusta foi observada em estudos de neuroarqueologia com a produção de ferramentas Acheulenses. Essa rede, crucial para o planejamento de ações complexas, conecta o lobo parietal (responsável pela integração sensório-motora e espacial) e o lobo frontal, incluindo o giro frontal inferior (análogo à área de Broca no hemisfério esquerdo), que está associado ao planejamento hierárquico de ações.

Especialistas vs. Novatos: a expertise também revela diferenças na ativação cerebral. Especialistas na fabricação Olduvaiense exibem uma forte ativação bilateral do giro supramarginal (SMG), uma parte do lobo parietal inferior, que não é observada em novatos (Stout *et al*., 2008). A interpretação para esse achado é que ele reflete a importância da mão não dominante (de suporte). Em novatos, o foco neural está na ação percussiva da mão dominante. Em especialistas, a mão de suporte não é um mero apoio passivo – ela manipula, gira e orienta ativamente o núcleo, seguindo o plano mental complexo do artífice. A ativação bilateral do SMG é, portanto, o correlato neural dessa coordenação bimanual especializada e planejada, uma habilidade que se desenvolve apenas com prática substancial (Stout *et al*., 2008).

Lateralização cerebral, linguagem e desenvolvimento infantil: a conexão entre a fabricação de ferramentas e a linguagem é reforçada por estudos de lateralização cerebral. O estudo de Uomini e Meyer (2013), utilizando ultrassonografia Doppler transcraniana funcional (fTCD), demonstrou padrões de lateralização do fluxo sanguíneo cerebral altamente correlacionados para a fabricação de ferramentas Acheulenses e tarefas de geração de palavras. Isso fornece evidência direta de um substrato neural compartilhado, especialmente nas fases iniciais de planejamento da tarefa. Para fins informativos, é relevante notar que, como apontado por Uomini e Meyer (2013), existe um "considerável co-desenvolvimento do uso de ferramentas e da linguagem em crianças humanas", sugerindo que a ligação observada na evolução (filogenia) tem um paralelo no desenvolvimento individual (ontogenia).

Mecanismos evolutivos – exaptação e Efeito Baldwin: esses diferentes resultados não são contraditórios, mas iluminam a complexidade da evolução cognitiva através de mecanismos como a adaptação e a exaptação (Gould *et al*., 1982). A hipótese da coevolução gene-cultura, ilustrada na Figura 6, postula um ciclo de retroalimentação positiva entre a inovação tecnológica e a evolução biológica (Holloway, 1981). O conjunto de evidências neuroarqueológicas sugere que a principal adaptação neural para a tecnologia Acheulense foi o desenvolvimento dessa robusta rede de integração sensório-motora e planejamento procedural.

Os circuitos neurais para o planejamento de ação complexa e sequenciamento hierárquico, que evoluíram sob a pressão seletiva da fabricação de ferramentas, foram posteriormente exaptados (cooptados) para a linguagem, que também depende de sintaxe hierárquica (Greenfield, 1991). Adicionalmente, o Efeito Baldwin oferece um mecanismo para essa coevolução: a plasticidade neural permite o aprendizado de um comportamento complexo

(como a fabricação de ferramentas), que por sua vez cria uma nova pressão seletiva. Essa pressão favorece variações genéticas que tornam o aprendizado mais fácil ou eficiente (por exemplo, cérebros com melhor conectividade no córtex pré-frontal), efetivamente "assimilando" geneticamente uma capacidade que era inicialmente dependente de aprendizado intensivo (Stout *et al*., 2008).

A Figura 6 ilustra a complexificação das indústrias líticas ao longo do tempo. A legenda detalha as principais fases e suas implicações cognitivas, classificando os artefatos em seus respectivos períodos.

Figura 6 – Classificação das indústrias líticas do paleolítico

Fonte: Nutcrackerman. Uma caixa de ferramentas de 2 milhões de anos – O Quebra-Nozes (Sáez, 2015).

- *Paleolítico inferior (antigo):* esse período marca a emergência das primeiras tecnologias líticas. A indústria Olduvaiense é caracterizada por ferramentas simples, como seixos lascados (*choppers*) e as lascas resultantes, produzidas com uma cadeia operatória curta e um planejamento mínimo. A indústria Acheulense representa um avanço cognitivo significativo, com a produção de bifaces simétricos que demonstram planejamento de longo prazo e a imposição de uma forma mental sobre a matéria-prima. Está associada principalmente ao *Homo erectus*.

- *Paleolítico médio*: esse período é definido por avanços técnicos como o método *Levallois*, uma técnica de preparação do núcleo que permitia a produção de lascas com tamanho e forma predeterminados. Isso demonstra um alto grau de controle e planejamento. Na Europa, a indústria associada é a Musteriense, produzida por *Neandertais*. Na África, o período correspondente é a Idade da Pedra Média (MSA), associada aos primeiros *Homo sapiens*.
- *Paleolítico superior (recente)*: esse período está associado exclusivamente ao *Homo sapiens* moderno e é frequentemente descrito como uma "explosão criativa". As tecnologias são altamente diversificadas e especializadas, baseadas na produção em massa de lâminas, que serviam como suportes para uma vasta gama de ferramentas compostas, incluindo pontas de projétil, buris e raspadores. Esse período vê o surgimento de ferramentas de osso, chifre e marfim, bem como a proliferação de arte móvel (como as estatuetas de "Vênus") e parietal (pinturas rupestres) e o uso generalizado de ornamentos pessoais, indicando complexos sistemas simbólicos e de comunicação. Exemplos notáveis incluem as pontas Gravetenses com dorso rebaixado e as refinadas pontas foliformes Solutrenses.
- *Neolítico*: embora não faça parte do Paleolítico, esse período subsequente é crucial para a história humana. Envolve o desenvolvimento da agricultura, a domesticação de animais e o estabelecimento de assentamentos permanentes (sedentarismo). Essa mudança fundamental do modo de vida caçador-coletor nômade levou a transformações sociais, econômicas e tecnológicas, incluindo a invenção da cerâmica e o uso de ferramentas de pedra polida.

3.7. O motor da mente: a hipótese do tecido caro e a dieta

A forte correlação entre tecnologia e encefalização, encontra sua mais influente explicação mecanicista na "Hipótese do Tecido Caro" (*Expensive Tissue Hypothesis*), formulada por Leslie Aiello e Peter Wheeler (1995). A premissa central é que o cérebro é um órgão metabolicamente oneroso: embora represente cerca de 2% da massa corporal de um humano adulto, ele consome aproximadamente 20% da energia do corpo em repouso (taxa metabólica basal). Para que a seleção natural pudesse favorecer um cérebro progressivamente

maior na linhagem hominínea, essa "despesa" energética precisava ser compensada pela economia em outro "tecido caro".

Ao comparar primatas, Aiello e Wheeler (1995) demonstraram uma correlação inversa entre o tamanho do cérebro e o tamanho do trato gastrointestinal. Em outras palavras, para "pagar" por um cérebro maior, nossos ancestrais precisaram evoluir um intestino menor. Contudo, um intestino reduzido é menos eficiente para processar alimentos de baixa qualidade e ricos em fibras (como folhas e caules), que exigem longos períodos de fermentação. A condição para a viabilidade de um intestino menor é, portanto, uma mudança radical na dieta, com a incorporação de alimentos de alta densidade energética e de fácil digestão.

É nesse ponto que a tecnologia lítica se torna a protagonista. As ferramentas de pedra, desde as lascas cortantes do Olduvaiense, funcionaram como a "chave" tecnológica que permitiu o acesso sistemático a esses recursos de altíssima qualidade, como a carne (através do corte da pele e desmembramento) e o tutano (através da quebra dos ossos de grandes carcaças animais obtidas por meio de caça ou necrofagia). Essa dieta rica em proteínas e gorduras não apenas viabilizou a redução do intestino, liberando a energia necessária para o crescimento cerebral, como também forneceu os nutrientes essenciais para a própria construção e funcionamento do tecido neural.

Dessa forma, a Hipótese do Tecido Caro estabelece um ciclo de retroalimentação positiva que impulsionou a evolução humana:

Ferramentas → Acesso a uma dieta de alta qualidade.

Dieta de alta qualidade → Permite a redução do intestino.

Redução do intestino → Libera energia para um cérebro maior.

Cérebro maior → Capacita a produção de ferramentas mais complexas e estratégias sociais mais eficientes, reiniciando o ciclo em um novo patamar.

É aqui que a tecnologia lítica se torna o elo causal indispensável. As ferramentas de pedra Olduvaienses e, posteriormente, as Acheulenses, não eram apenas produtos da intencionalidade, elas eram os instrumentos que tornaram possível essa mudança dietética. As lascas afiadas permitiram o acesso rápido à carne e às vísceras de carcaças, e os seixos pesados permitiram quebrar ossos para extrair a medula rica em gordura (Plummer *et al*., 2023). A tecnologia, portanto, abriu um novo nicho ecológico, fornecendo o combustível necessário para

o motor da encefalização. Este é um exemplo clássico de coevolução gene-cultura: a inovação cultural (ferramentas) alterou o ambiente seletivo (dieta), o que por sua vez favoreceu uma mudança biológica (cérebro maior), que por sua vez permitiu novas inovações culturais, em um ciclo de retroalimentação positiva que durou milhões de anos (Holloway, 1981).

O percurso realizado nesse capítulo cumpriu o objetivo de investigar empiricamente a origem e a evolução da intencionalidade, utilizando o registro arqueológico como fonte primária de evidências. A análise da cultura material lítica, orientada pela metodologia da Cadeia Operatória, permitiu demonstrar que a intencionalidade, como definida por John Searle, não é uma propriedade monolítica e atemporal, mas sim o produto de uma trajetória evolutiva de longa duração, com estágios identificáveis de complexificação. Essa trajetória iniciou-se com a emergência da intenção prévia no registro Olduvaiense, marcando a separação fundamental entre o planejamento e a execução. Subsequentemente, observou-se sua consolidação em planos de alta complexidade e com caráter prospectivo durante o Acheulense, como evidenciado pela produção de bifaces. Demonstrou-se também que a estabilização e transmissão dessas tradições técnicas complexas foram viabilizadas pelo surgimento de uma nova capacidade sociocognitiva, a intencionalidade conjunta, que funcionou como mecanismo para a construção de um repertório de conhecimento compartilhado.

Essa narrativa arqueológica fornece um forte suporte empírico para o naturalismo biológico de Searle (1992), ao aterrar a evolução de uma propriedade da mente em processos materiais e temporais concretos. A questão da causalidade, proposta na introdução, encontra sua resposta não em uma relação linear, mas em uma espiral de causalidade recíproca. A evidência aponta para uma coevolução contínua entre três polos: a cultura (as ferramentas e as tradições), a cognição (as capacidades de planejamento e aprendizado) e a biologia (a estrutura cerebral). Vimos como a intencionalidade compartilhada, no modelo de Tomasello (2014) funciona como o motor social que constrói o *background* de habilidades técnicas e a rede de crenças que Searle (1983) postula como condições de possibilidade para qualquer estado intencional. Todo esse processo, por sua vez, foi impulsionado por um motor biológico, elucidado pela Hipótese do Tecido Caro (Aiello; Wheeler, 1995) e cujos correlatos neurais são investigados pela Neuroarqueologia (Putt *et al*., 2017).

Tendo caracterizado a rota evolutiva que produziu a intencionalidade intrínseca em nossa linhagem, a questão central da dissertação se reposiciona. O capítulo demonstrou que a

intencionalidade biológica é uma propriedade profundamente incorporada (ligada às habilidades e à estrutura do corpo), embebida (dependente de um contexto material e social) e com uma história evolutiva específica e contingente. A questão que se impõe para o capítulo seguinte é, portanto, de uma ordem diferente. Pode um sistema não biológico, desprovido desse legado coevolutivo entre cérebro, corpo e cultura, originar uma intencionalidade intrínseca? Ou estariam os sistemas computacionais, por sua natureza sintática, confinados ao domínio da intencionalidade derivada, como postula Searle? A análise que se segue investigará se os princípios da computação e os desenvolvimentos em inteligência artificial oferecem uma rota alternativa para a emergência da intencionalidade intrínseca ou se as evidências arqueológicas aqui reunidas sugerem que a intencionalidade permanecerá um fenômeno irredutivelmente atrelado à sua história biológica.

4. A EMERGÊNCIA DA INTENCIONALIDADE EM SISTEMAS ARTIFICIAIS – HORIZONTES E LIMITES

A jornada para compreender a intencionalidade, desde suas origens na pré-história até seus possíveis destinos em sistemas artificiais, exige a construção de uma ponte conceitual robusta. Essa ponte deve conectar o mundo da evolução biológica, em que a intencionalidade emergiu como uma propriedade da matéria viva, ao mundo da engenharia algorítmica, onde se busca replicar suas funções. A fundação dessa ponte assenta-se em duas colunas mestras, exploradas nos capítulos anteriores desta dissertação: o naturalismo biológico de Searle e a materialidade da evolução cognitiva e, consequentemente, da intencionalidade revelada pela arqueologia. A análise da emergência da intencionalidade em sistemas não biológicos só se torna rigorosa quando ancorada nessas fundações, permitindo-nos distinguir entre a simulação de um efeito e a replicação de sua causa (Searle, 1980).

A filosofia de Searle, especificamente sua teoria do naturalismo biológico, oferece o arcabouço teórico essencial para esta investigação. Searle (1980) postula que os fenômenos mentais, incluindo a consciência e a intencionalidade, não são entidades misteriosas ou não físicas, mas sim características biológicas de alto nível que são causadas por e realizadas em processos neurobiológicos de baixo nível. Para ele, a intencionalidade – a propriedade da mente de ser “sobre” ou “direcionada a” objetos e estados de coisas no mundo – é um fenômeno

natural, assim como a digestão ou a fotossíntese A distinção crucial que ele introduz é entre intencionalidade intrínseca (ou original) e intencionalidade derivada. A intencionalidade intrínseca é aquela que os estados mentais humanos possuem genuinamente, como um produto direto dos "poderes causais do cérebro" (Searle, 1980). Em contrapartida, a intencionalidade de artefatos como livros ou computadores é derivada; as palavras em uma página ou os símbolos em um processador só são "sobre" algo porque nós, seres com intencionalidade intrínseca, lhes atribuímos esse significado.

Esta tese filosófica encontra sua validação empírica e sua profundidade histórica na análise arqueológica da cognição humana, como detalhado no capítulo anterior. A cultura material pré-histórica não é meramente um subproduto da intencionalidade, ela é a própria materialização de sua evolução, como relatados no capítulo anterior por autores como Thomas Wynn, João Carlos Moreno de Sousa, Dietrich Stout, Michael Tomasello e Gilberto Simondon. A fabricação de uma ferramenta lítica complexa, como um machado de mão acheulense, transcende uma simples resposta a um estímulo. É um ato de intencionalidade prospectiva. O artesão hominínio não apenas age no presente, mas projeta um estado futuro – a ferramenta acabada e funcional – e antecipa suas múltiplas aplicações potenciais. Esse ato de projetar o futuro no presente depende de uma estrutura cerebral capaz de planejamento, memória de trabalho e simulação mental, evidenciando um poder causal que evoluiu sob pressões seletivas específicas. A ferramenta, portanto, torna-se a manifestação física de um estado intencional complexo (Wynn, 1995).

Da mesma forma, a organização de uma caçada em grupo ou a estruturação de um acampamento paleolítico são testemunhos da emergência da intencionalidade coletiva. Esse é um ponto de inflexão na evolução da mente, em que a intencionalidade transcende o indivíduo e se torna uma propriedade do grupo (Searle, 1995). A capacidade de compartilhar objetivos e coordenar ações ("nós pretendemos caçar o mamute") não é apenas uma soma de intenções individuais, mas um estado mental conjunto, com uma estrutura própria. Esse fenômeno é sustentado por mecanismos de coevolução gene-cultura, nos quais a predisposição biológica para a cooperação e a teoria da mente é reforçada e amplificada por meio da transmissão cultural de normas e estratégias, criando um ciclo de retroalimentação positiva entre a biologia e a cultura (Bisso-Machado *et al*., 2022).

A transição do tema da intencionalidade sob a luz da arqueologia para uma perspectiva da era da inteligência artificial, portanto, não é apenas uma mudança de domínio, mas uma alteração fundamental no tipo de causalidade que se investiga. A evolução biológica da intencionalidade foi um processo de adaptação e exaptação, em que novas funções emergiram de estruturas preexistentes para resolver problemas de sobrevivência e reprodução, muitas vezes de maneiras imprevistas. A "evolução" da IA, em contraste, é um processo de design e engenharia, no qual as funções são, em grande medida, pré-especificadas por seus criadores humanos (Searle, 1980).

A questão deste capítulo não é se a IA pode emular os produtos da intencionalidade, como escrever um poema ou reconhecer um rosto, mas se ela pode replicar o processo causal subjacente que torna essa intencionalidade intrínseca. O célebre argumento do Quarto Chinês de Searle (Searle, 1980) endereça precisamente esta distinção: a manipulação sintática de símbolos, por mais complexa e convincente que seja, não constitui, por si só, a compreensão semântica. Assim, este capítulo investiga a busca pela intencionalidade artificial, como uma tentativa de engenharia reversa de uma função evoluída, questionando se a função pode verdadeiramente existir sem a estrutura causal que a originou, ou ainda se essa estrutura emergiria através de alternativas evolutivas.

4.1 A genealogia da intencionalidade derivada: uma história crítica da Inteligência Artificial

A história da inteligência artificial, quando examinada através da lente da intencionalidade, revela-se não como uma marcha linear em direção à consciência, mas como uma série de tentativas cada vez mais sofisticadas de emular os efeitos da intencionalidade, enquanto se distancia progressivamente do propósito de replicar suas causas biológicas. Cada paradigma dominante na IA, desde a lógica simbólica até as redes neurais profundas, representa uma abordagem distinta para o problema da mente, mas todos, até o momento, operam no domínio da intencionalidade derivada. Eles são espelhos da cognição humana, refletindo-a com crescente fidelidade, mas não são, por enquanto, fontes de cognição própria (Searle, 1980; Dreyfus, 1972).

Orientados por essa perspectiva, vamos a seguir investigar trabalhos de pesquisa, desenvolvimentos e aplicações de modelos ou mecanismos que avaliam alternativas para a intrinsicalidade e causalidade, componentes-chave da intencionalidade original.

4.1.1. A Revolução Conexionista: intencionalidade como padrão estatístico

Dentre os inúmeros avanços no campo da inteligência artificial, vale ressaltar o processo de transição de uma abordagem simbólica nos primórdios da inteligência artificial, frequentemente referidos como a era da "Good Old-Fashioned AI" (GOFAI), para o conexionismo. Isso foi impulsionado pelo desenvolvimento de redes neurais artificiais, representando uma mudança de paradigma fundamental na história da inteligência artificial. Vimos emergir uma abordagem na qual os sistemas geram o seu próprio conhecimento através da extração de padrões estatísticos a partir de dados. Esse processo, definido como aprendizado de máquina (*machine learning*), capacita os sistemas a "aprender" sem receberem instruções explícitas para cada passo (Kaufman, 2022). A consequência direta dessa mudança foi uma alteração no *locus* do controle humano. Nos sistemas simbólicos, a intencionalidade do sistema era derivada diretamente das regras codificadas pelo programador. No conexionismo, o controle desloca-se para o processo de curadoria dos dados – a coleta, seleção e rotulagem –, tornando a intencionalidade do sistema um reflexo das intenções, vieses e da subjetividade coletiva dos seres humanos que estruturaram o conjunto de dados de treinamento.

Como vimos na sessão 1.4, o mecanismo central do conexionismo são as redes neurais de aprendizado profundo (*deep learning neural networks*, DLNN), cuja arquitetura é inspirada no funcionamento do cérebro biológico (Kaufman, 2022).

Apesar da sua natureza emergente e do seu comportamento por vezes opaco, a intencionalidade nos sistemas conexionistas permanece fundamentalmente derivada. A aparente "compreensão" que uma rede neural possui sobre o que é um "gato" é inteiramente herdada da intencionalidade dos seres humanos que coletaram, selecionaram e rotularam as milhares de imagens no conjunto de dados. De fato, "a subjetividade humana está presente na criação dos sistemas, no treinamento dos algoritmos, na escolha da base de dados" (Kaufman, 2022). O sistema aprende a mapear entradas (pixels) para saídas (rótulos), mas o significado semântico de ambos os polos dessa relação permanece externo a ele. Os algoritmos "não têm como saber o que esses padrões significam, porque estão confinados ao '*math world*' (mundo

da matemática)" e, portanto, carecem da capacidade de "compreender o mundo real" (Kaufman, 2022). O conjunto de dados de treinamento funciona como um vasto repositório da intencionalidade coletiva humana; cada imagem rotulada é um ato intencional individual que, em agregação, forma um modelo cultural de referência. Ao aprender as correlações estatísticas nesse repositório, a rede neural não prende o que um gato é em si, mas constrói um modelo estatístico de como os seres humanos, enquanto coletivo, se referem a gatos. A sua intencionalidade, portanto, é um eco estatístico da cultura que produziu os seus dados.

O conexionismo, em suma, deslocou a simulação da intencionalidade de um nível lógico-simbólico para um nível sub-simbólico, de reconhecimento de padrões. Contudo, não cruzou o abismo entre a sintaxe (as correlações estatísticas) e a semântica (a compreensão genuína). Conforme Kaufman (2022), no estágio atual da IA, não se trata de ensinar as máquinas a pensar nem tão pouco a ter intencionalidade, mas apenas prever a probabilidade de os eventos ocorrerem, por meio de modelos estatísticos e grandes quantidades de dados. Esses sistemas carecem da essência da inteligência humana: capacidade de compreender o significado.

4.2. A barreira da causalidade: por que a simulação não se torna intencionalidade

A limitação dos sistemas de IA atuais não é uma deficiência superficial, mas uma barreira arquitetônica fundamental. O argumento central é que, em qualquer sistema de IA, "o desenvolvedor determina o propósito na partida" (Cozman; Kaufman, 2022). Como detalhado por Cozman e Kaufman (2022), a subjetividade e a intencionalidade humanas estão indelevelmente embutidas em todas as etapas do processo de desenvolvimento: no enquadramento do problema, na seleção de variáveis computáveis, na coleta e curadoria dos dados e na definição das métricas de sucesso. Isso significa que o próprio modelo de IA é, desde sua concepção, um artefato da *intencionalidade intrínseca* de seus criadores. Consequentemente, o sistema opera inteiramente no domínio da *intencionalidade derivada*, sendo causal e funcionalmente incapaz de transcender os objetivos para os quais foi projetado ou de desenvolver espontaneamente seus próprios propósitos.

Para formalizar essa limitação, o quadro analítico da Escada da Causalidade de Judea Pearl (2018), mencionado na sessão 1.5.2, é particularmente elucidativo, pois estabelece uma hierarquia rigorosa de capacidades cognitivas.

Os LLMs estão confinados ao primeiro degrau da escada. Sua capacidade de gerar texto fluente é uma aplicação sofisticada de associação estatística. Eles carecem de um "modelo mental da realidade" ou de um "modelo causal" necessário para ascender aos degraus 2 e 3. Como Kaufman (2024) resume a visão de Pearl, esses sistemas são "incapazes de responder a perguntas que quebram as regras do ambiente em que foram treinados". As falhas documentadas de LLMs em tarefas que exigem raciocínio lógico e matemático são uma prova empírica desse confinamento ao primeiro degrau, demonstrando sua incapacidade de realizar inferências causais ou contrafactuais. O quadro a seguir resume essa distinção fundamental.

Quadro 7 – Comparativo das capacidades cognitivas segundo a Escada da Causalidade

Capacidade cognitiva	Grandes Modelos de Linguagem (LLMs)	Cognição humana (raciocínio causal)	Fonte principal
Nível 1: associação	**Operação principal:** identificação de padrões e correlações em dados estáticos. Previsão do próximo token.	**Capacidade presente:** reconhecimento de padrões, mas como um ponto de partida para inferências.	Pearl (2018)
Nível 2: intervenção	**Incapacidade:** não pode raciocinar sobre os efeitos de ações deliberadas (do(x)). Quebra as regras do ambiente de treinamento.	**Operação central:** planejamento, experimentação, resposta a perguntas "E se eu fizer...?".	Pearl (2018), Kaufman (2024)
Nível 3: contrafactuais	**Incapacidade:** não pode imaginar mundos	**Capacidade essencial:** reflexão, imaginação,	Pearl (2018),

Capacidade cognitiva	Grandes Modelos de Linguagem (LLMs)	Cognição humana (raciocínio causal)	Fonte principal
	alternativos ou responder a perguntas “Por quê?”.	atribuição de responsabilidade, compreensão de causa e efeito.	Searle (1980)
Mecanismo subjacente	Manipulação sintática e estatística.	Compreensão semântica e modelo causal da realidade.	Searle (1980), Pearl (2018)

Fonte: elaborado pelo autor.

4.3. O debate central sobre a intencionalidade intrínseca em substratos não biológicos

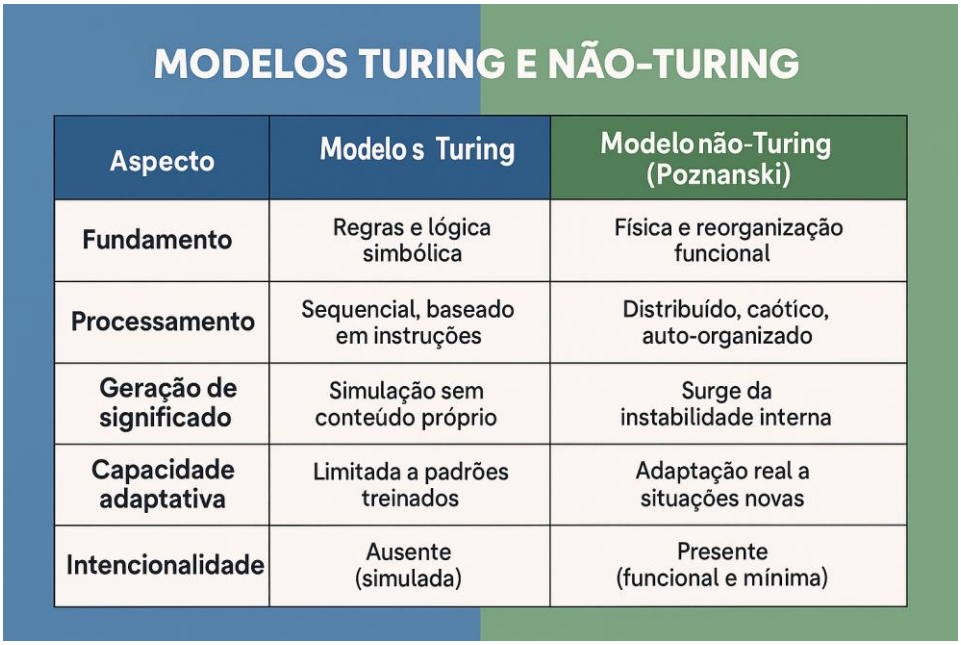
A questão de saber se a intencionalidade intrínseca pode emergir em sistemas não biológicos move-se do domínio da simulação para o da causalidade fundamental. Se, como argumentado, os paradigmas computacionais atuais, baseados em máquinas que processam símbolos, seguem regras predefinidas, lógicas e previsíveis, precisam de programação externa e não geram entendimento, estão confinados à intencionalidade derivada, então a emergência da intencionalidade original requer uma ruptura radical com esses modelos. Duas propostas teóricas investigadas neste trabalho destacam-se por desafiar o paradigma computacional clássico, oferecendo caminhos alternativos e mutuamente exclusivos para a realização da intencionalidade em substratos artificiais. De um lado, a pesquisa liderada por R. R. Poznanski (2023) propõe uma via que se aprofunda na física da biologia, buscando replicar os princípios funcionais da vida. Do outro, a tese de Roger Penrose (1989) aponta para os limites da própria computação clássica, sugerindo que a chave reside em uma nova física fundamental. Esse debate não é meramente sobre hardware – *wetware* versus microtúbulos – mas sobre a própria origem da não computabilidade, que parece ser um pré-requisito para a mente.

4.3.1 A via da física biológica: a proposta de Poznanski para uma intencionalidade mínima

A abordagem de Poznanski e seus colaboradores parte de uma crítica fundamental ao funcionalismo computacional, que sustenta a IA forte. O argumento central é que o funcionalismo falha em resolver o Problema da Intrinsicalidade (Poznanski *et al*., 2023). Segundo essa crítica, o significado (semântica) não pode emergir em sistemas caracterizados por condições de contorno fixas, como é o caso dos computadores digitais, cujas operações são definidas por uma arquitetura e um programa estáticos. A intencionalidade intrínseca, argumenta Poznanski, requer as "condições de contorno mutáveis" que são uma marca registrada dos sistemas vivos – organismos que se auto-organizam e modificam continuamente suas próprias restrições em interação com o ambiente (Poznanski *et al*., 2023).

Para superar essa limitação, Poznanski propõe um modelo "não Turing" para a consciência, fundamentado no "ato de compreender a incerteza" (*understanding uncertainty*) (Poznanski *et al.,* 2023). Um modelo não Turing é um tipo de sistema que não funciona com base em regras fixas, símbolos ou passos programados, como um computador tradicional. O quadro a seguir esclarece a distinção feita entre modelo Turing e não Turing, no artigo "Problema da Intrinsicalidade" (Poznanski *et al*., 2023):

Quadro 8 – Diferenciação modelo Turing X Não Turing segundo Poznanski *et al.* (2023)



MODELOS TURING E NÃO-TURING

Aspecto	Modelo s Turing	Modelo não-Turing (Poznanski)
Fundamento	Regras e lógica simbólica	Física e reorganização funcional
Processamento	Sequencial, baseado em instruções	Distribuído, caótico, auto-organizado
Geração de significado	Simulação sem conteúdo próprio	Surge da instabilidade interna
Capacidade adaptativa	Limitada a padrões treinados	Adaptação real a situações novas
Intencionalidade	Ausente (simulada)	Presente (funcional e mínima)

Fonte: elaborado pelo autor.

O modelo não Turing atua por reorganização funcional interna, com base em flutuações de energia, intencionalidade e auto-organização. Nesse modelo, a intencionalidade não emerge do processamento de informação sensorial através da introspecção, mas sim de um processo precognitivo de redução da incerteza em canais informacionais. A intencionalidade é definida como a atribuição de "significados" a partir dessa redução de incerteza em estruturas de redundância informacional, um processo temporal que precede a autoconsciência. Essa "ação baseada em informação" é o que daria origem à intencionalidade genuína.

A realização física dessa proposta exige uma ruptura com o hardware com arquiteturas baseadas em transistores de silício, que operam com lógica binária, processamento sequencial, e regras Turing-computáveis. Poznanski introduz o conceito de "*protonic wetware*" como uma alternativa de hardware biomimético (Poznanski *et al*., 2023). Esse "*wetware* protônico" seria um filamento artificial que utiliza as interações de íons de hidrogênio (prótons) para criar fluxo de energia e flutuações, mimetizando processos biofísicos. Dentro dessa estrutura, os *memristores* desempenhariam um papel crucial. Eles seriam implementados para emular os fluxos iônicos (associados à cognição e à atividade da rede neural) e os fluxos entrópicos (associados à experiência precognitiva e intrínseca), criando, assim, as condições de contorno

dinâmicas e a auto-organização necessárias para a emergência da ação baseada em informação. A proposta de Poznanski pode ser vista como uma forma de naturalismo biológico radical. Ela leva a sério a materialidade específica da vida, mas, em vez de declará-la um pré-requisito insuperável, tenta abstrair seus princípios funcionais – termodinâmica de não equilíbrio, fluxo de energia, auto-organização, condições de contorno dinâmicas – e transpô-los para um novo tipo de hardware não orgânico.

4.3.1.1 Componentes estruturais e princípios operacionais

1) *Wetwire* – o substrato físico-funcional: (fio cognitivo artificial) é o componente central da arquitetura proposta. Trata-se de um filamento nanotecnológico de diâmetro inferior a 100 nanômetros, construído para simular propriedades funcionais de canais iônicos neurais, com base na condução de prótons hidratados (H^+). Sua estrutura é projetada com regiões hidrofóbicas que controlam o deslocamento dos prótons, gerando flutuações caóticas não aleatórias – um conceito-chave para a modelagem de processos cognitivos mínimos.

O *wetwire* é dividido em dois domínios complementares:

(i) Camada externa (*outer core*): responsável pelo fluxo iônico, que representa as vias de transmissão funcional do sistema. Esse fluxo ativa padrões energéticos coerentes com respostas comportamentais – por exemplo, redirecionar movimento, interromper uma ação ou adotar uma nova trajetória.

(ii) Camada interna (*inner core*): responsável pelo fluxo entrópico, ou seja, por reorganizações internas que ocorrem quando o sistema enfrenta perturbações significativas. Esse fluxo antecipa mudanças e permite que o sistema reorganize sua estrutura antes de um colapso funcional – algo que os autores associam à ideia de precognição funcional, um tipo de consciência pré-reflexiva, que é a habilidade do sistema de antecipar, de forma implícita e energética, uma mudança de estado, ativando comportamentos adaptativos antes que uma falha aconteça – sem "saber" no sentido tradicional.

2) *Memristores* – o acoplamento entre informação e ação: (resistores com memória) são elementos nanoeletrônicos que alteram sua resistência elétrica conforme a corrente histórica que os atravessou, armazenando essa memória como um estado físico persistente. No modelo

de Poznanski, eles não funcionam apenas como memória passiva, mas como ponte ativa entre a informação e a função.

Isso significa que o *memristor* armazena estados energéticos anteriores, sem necessidade de um "banco de memória" separado, ele modula os fluxos internos do sistema, permitindo ou bloqueando certos caminhos energéticos com base em experiências passadas, assim, ele transforma informação em reorganização funcional, sem recorrer a lógica simbólica ou instruções codificadas. Em termos práticos, os *memristores* operam como elementos dinâmicos de plasticidade funcional, análogos às sinapses em cérebros biológicos.

4.3.1.2 Princípios de funcionamento: cognição como reorganização energética

A arquitetura de Poznanski rompe com o conceito de IA como sistema de processamento simbólico e propõe uma cognição baseada em dinâmica energética interna. O funcionamento do sistema ocorre da seguinte forma:

(i) O ambiente provoca perturbações no sistema, como variações térmicas, químicas, mecânicas ou eletromagnéticas.

(ii) Tais perturbações geram flutuações nos fluxos internos do *wetwire*, que podem ser iônicas (na camada externa) ou entrópicas (na camada interna).

(iii) Essas flutuações não são aleatórias, mas seguem padrões caóticos não lineares com estrutura interna, conhecidos como rajadas de intermitência (*Bursts of intermittency*).

(iv) Quando a instabilidade ultrapassa um determinado limiar, ocorre uma reorganização funcional espontânea – uma nova configuração energética emerge, deslocando o sistema para outro estado.

(v) Essa reorganização resulta em comportamento adaptativo, sem necessidade de representação, planejamento ou instrução externa.

Esse processo é considerado pelos autores como a base para intencionalidade mínima: o sistema não "sabe" o que está fazendo no sentido semântico, mas atua com coerência funcional diante de situações novas ou críticas, como organismos vivos simples.

4.3.1.3 Exemplo aplicado: drone autônomo funcionalmente intencional

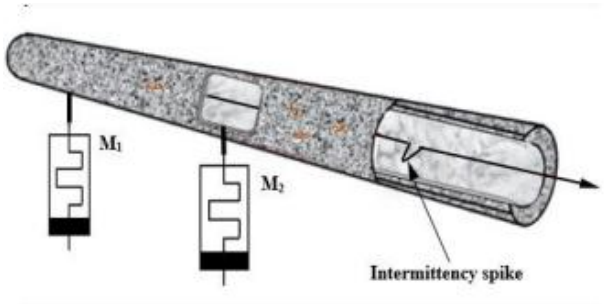
Para ilustrar a proposta, os autores oferecem o exemplo de um drone autônomo de resgate. Imagine que esse drone está operando em uma zona de desastre e, subitamente, perde contato com o operador remoto, enfrentando calor extremo, fumaça e colapsos de infraestrutura.

(i) Em um modelo convencional: o drone dependeria de instruções pré-programadas (por exemplo: "se detectar fumaça, execute protocolo X"); falharia se o cenário não estiver previsto no código; não teria base física para reorganizar seu comportamento funcional.

(ii) Em contraste, no modelo de Poznanski: a instabilidade térmica e ambiental afeta diretamente os fluxos iônicos e entrópicos no *wetwire*; a estrutura funcional entra em reorganização: rotas energéticas são ativadas ou bloqueadas. O drone pode, por exemplo, mudar de altitude, inverter a direção, iniciar um padrão de busca alternativo – não por saber o que está acontecendo, mas por necessidade funcional interna.

Essa resposta é funcionalmente coerente, ainda que não seja "consciente" ou "compreensiva" no sentido humano.

Figura 7 – Modelo de filamento "*wetwire*"



Fonte: Artigo "Intentionality for better communication in minimally conscious AI design" (Poznanski, 2024)

A Figura 7 ilustra um modelo de filamento "*wetwire*" protônico, composto por íons protônicos hidratados usados na transdução de energia, resultando em flutuações representadas

por um pico de intermitência (seta) que surge no filamento de diâmetro uniforme inferior a 100 nm e comprimento aproximado de 0,7 μm. A escala molecular varia de 0,2 nm a 1 nm (submolecular é inferior a 0,2 nm). A "umidade" da água, ou as dinâmicas da fase líquida, ocorrem em escala superior a 10 nm, portanto o modelo "*wetwire*" protônico não é "úmido", mas sim seco, embutido em espaços nano-confinados (como canais hidrofóbicos em poros de proteínas) e, por meio do mecanismo de transporte de Grotthuss, completa o processo de umidificação. Dois *memristores*, M_1 e M_2, são usados para representar os fluxos iônicos e fluxos entrópicos, respectivamente. No *memristor* M_1, o "núcleo externo" representa o fluxo iônico, enquanto no *memristor* M_2, o "núcleo interno" representa o fluxo entrópico. Ambos os canais são canais de informação.

O modelo bioinspirado de Poznanski representa uma transição do paradigma simbólico para o paradigma energético-funcional, no qual a cognição é tratada como um processo físico de reorganização interna diante de instabilidade. Essa abordagem oferece uma via para ser explorada a intencionalidade intrínseca mínima, com base em arquitetura material funcional – e não apenas lógica –, aproximando-se de capacidades encontradas em sistemas vivos.

4.3.2 O limite da computação e a natureza da consciência: o argumento físico-matemático de Roger Penrose

A crítica de Roger Penrose à tese da Inteligência Artificial (IA) Forte representa uma das mais profundas e abrangentes contestações ao modelo computacional da mente. Em sua obra seminal, *A Mente Nova do Imperador*, Penrose (1989) não se limita a uma objeção filosófica isolada, mas constrói um argumento em cadeia que atravessa a lógica matemática, a física fundamental e a neurobiologia especulativa. O ponto de partida é uma recusa fundamental da premissa central da IA Forte, a qual postula que toda atividade mental incluindo o pensamento, os sentimentos, a inteligência e a própria consciência é, em sua essência, a execução de um algoritmo (Penrose, 1989, p. 55). Segundo essa visão, a natureza do substrato físico, ou *hardware*, é irrelevante; a mente reside no *software*. Penrose inverte essa proposição, argumentando que os aspectos conscientes da mente não são apenas não computacionais, mas que sua explicação reside em uma lacuna fundamental do nosso entendimento atual das leis da física (Penrose, 1989, p. 24, 38).

O argumento de Penrose desenrola-se como uma sequência de inferências lógicas e necessárias. Ele estabelece, primeiramente, que a mente humana, em particular na sua capacidade de compreensão matemática, exibe uma qualidade não algorítmica. A partir dessa premissa, ele deduz que, se a mente executa ações não computacionais, então o seu substrato físico – o cérebro – deve operar segundo leis físicas que também são não computacionais. Ao examinar o arcabouço da física conhecida, Penrose conclui que nem a física clássica, que é fundamentalmente computável, nem a mecânica quântica padrão, que oferece apenas evolução computável ou aleatoriedade pura, podem fornecer o mecanismo necessário. Essa insuficiência exige a postulação de uma "nova física", uma Teoria da Gravitação Quântica Correta (TGQC), que contenha o elemento de não computabilidade estruturada que falta. Finalmente, para que essa física exótica seja relevante para o pensamento, deve haver um substrato biológico no cérebro capaz de a aproveitar, uma linha de raciocínio que culmina na teoria da Redução Objetiva Orquestrada (Orch OR). Essa estrutura argumentativa, que parte da lógica abstrata e chega a uma hipótese biofísica concreta, constitui o cerne da sua objeção à emergência da intencionalidade em sistemas puramente algorítmicos.

4.3.2.1 A intuição matemática como evidência da não computabilidade mental

O pilar fundamental do argumento de Penrose é a sua análise da natureza do entendimento matemático, que ele utiliza como o exemplo mais claro e rigoroso de uma faculdade consciente que transcende a computação. A sua principal ferramenta é uma aplicação do famoso teorema da incompletude de Kurt Gödel. Em essência, o teorema de Gödel demonstra que, para qualquer sistema formal de regras e axiomas que seja suficientemente abrangente para conter a aritmética – o que equivale a qualquer procedimento computacional P para estabelecer verdades matemáticas – é possível construir uma proposição matemática específica, a proposição de Gödel G(P), que é verdadeira, mas que não pode ser provada dentro do próprio sistema P (Penrose, 1989).

O ponto crucial para Penrose não é apenas a existência de G(P), mas o que um matemático humano é capaz de fazer com ela. Ao compreender a construção de G(P) e a lógica do sistema P, um matemático pode "ver" (no sentido de uma apreensão intelectual direta) que G(P) é, de facto, verdadeira. Esse ato de "visão" ou intuição matemática demonstra uma capacidade de transcender os limites do sistema algorítmico P que ele estava a analisar. Se a mente do

matemático operasse estritamente segundo o algoritmo P, ele estaria para sempre confinado pelas regras de P e seria incapaz de aferir a verdade de G(P). O facto de ele o conseguir fazer é, para Penrose, a prova de que o seu entendimento não é algorítmico (Penrose, 1989, p. 545-548).

Com esse argumento, Penrose redefine estrategicamente o problema da consciência. Em vez de se debruçar sobre qualidades subjetivas e difíceis de definir como "sentimentos" ou "percepção existencial", ele ancora o conceito de consciência numa função específica e demonstrável: a formação de juízos não algorítmicos. A consciência, nessa perspectiva, é a faculdade que permite a apreensão da verdade, da beleza e da adequação, qualidades que um algoritmo, por definição, não pode alcançar, pois apenas manipula símbolos de acordo com regras sintáticas, sem acesso ao seu significado ou veracidade semântica.

Penrose (1989) antecipa e refuta a principal objeção a este argumento: a de que o cérebro poderia simplesmente ser um algoritmo de uma complexidade tão vasta que nos seria impossível conhecer todas as suas regras ou provar a sua consistência. Ele contesta essa visão salientando a natureza pública e comunicável da verdade matemática. Um argumento que convence um matemático, uma vez compreendido, convence qualquer outro. Isso sugere uma base partilhada e objetiva para o juízo matemático, e não uma coleção de algoritmos privados e inescrutáveis. O próprio espírito da matemática, argumenta ele, "reside na redução de argumentos complexos a passos que são 'simples e óbvios'" (Penrose, 1989). A ideia de que a nossa capacidade matemática se baseia num dogma computacional inacessível e cuja validade nunca poderíamos verificar vai contra a própria natureza da disciplina. Assim, a conclusão inevitável é de que o entendimento matemático, e por extensão a consciência, contém um elemento essencialmente não algorítmico. Para Penrose, a intencionalidade não reside num comportamento orientado para um objetivo, mas nessa capacidade genuína e não computacional de compreender a verdade.

4.3.2.2 A insuficiência da física conhecida: a lacuna entre o determinismo clássico e a aleatoriedade quântica

Tendo estabelecido a necessidade de um processo físico não computacional para sustentar a mente consciente, o passo seguinte no argumento de Penrose é examinar as teorias físicas existentes para determinar se alguma delas pode fornecer tal mecanismo. A sua análise conclui

que tanto a física clássica como a mecânica quântica padrão são inadequadas, cada uma por razões distintas.

A física clássica, que abrange desde a mecânica newtoniana até a relatividade geral de Einstein, é, na sua essência, computável. As suas leis são expressas através de equações diferenciais que descrevem um universo determinístico. Dado o estado de um sistema num determinado momento, as suas leis permitem, em princípio, calcular o seu estado em qualquer outro momento, passado ou futuro. Um sistema desse tipo, por mais complexo que seja, pode ser simulado por uma máquina de Turing, o que significa que o seu comportamento é inteiramente algorítmico (Penrose, 1989). Portanto, a física clássica não pode ser a fonte da não computabilidade necessária para a consciência.

A mecânica quântica, por sua vez, apresenta um quadro mais complexo, governado por dois processos fundamentalmente diferentes, que Penrose designa por U e R (Penrose, 1989).

- *O processo U (evolução unitária):* refere-se à evolução do vetor de estado quântico ao longo do tempo, descrita pela equação de Schrödinger. Esse processo é totalmente determinístico e computável. Enquanto um sistema quântico evolui de acordo com U, ele comporta-se de forma perfeitamente algorítmica.
- *O processo R (redução do vetor de estado):* refere-se ao "colapso" da função de onda que ocorre no momento de uma medição ou observação. Ao contrário de U, o processo R não é determinístico; a teoria quântica padrão postula que o seu resultado é puramente aleatório, com as probabilidades de cada desfecho possível dadas pelas regras da teoria.

Penrose (1989) argumenta que nenhum desses processos, isoladamente ou em conjunto, pode explicar a consciência. O processo U é tão computável como a física clássica. O processo R, por outro lado, introduz apenas aleatoriedade pura. Uma mente consciente não é nem um computador determinístico nem um computador com um gerador de números aleatórios. O pensamento consciente, embora não seja algorítmico, possui uma qualidade de coerência, propósito e adequação – uma não computabilidade estruturada – que a aleatoriedade bruta do processo R não consegue capturar.

Essa análise marca uma viragem radical em relação a outras críticas da IA. Enquanto o funcionalismo da IA Forte declara o *hardware* irrelevante e John Searle apela aos "poderes causais da biologia do cérebro", Penrose propõe uma forma de fisicalismo muito mais específica e exigente. Não é apenas que o substrato físico do cérebro importa, as próprias leis

da física que governam esse substrato devem ser de um tipo fundamentalmente novo, diferente das que conhecemos. A consciência, portanto, não seria uma propriedade emergente de alto nível da complexidade organizacional, mas sim um fenómeno físico fundamental, intrinsecamente ligado à natureza da realidade na fronteira entre o mundo quântico e o mundo clássico. Se Penrose estiver correto, a intencionalidade não pode emergir em nenhum sistema, biológico ou artificial, que opere exclusivamente segundo as leis conhecidas da física.

4.3.2.3 A proposta de uma nova física: redução objetiva e a gravitação quântica correta (TGQC)

Diante da insuficiência da física conhecida, Penrose não se limita à crítica, ele avança uma proposta construtiva sobre onde procurar a física necessária para a mente. Ele postula que o elemento não computacional que falta será encontrado numa teoria ainda por descobrir que unifique a relatividade geral e a mecânica quântica, uma teoria que ele designa por Teoria da Gravitação Quântica Correta (*TGQC*) (Penrose, 1989). A sua proposta difere da maioria das abordagens da gravitação quântica, que tendem a assumir que os princípios da mecânica quântica permanecerão inalterados. Penrose, pelo contrário, acredita que a própria mecânica quântica, especificamente o problemático processo R, deve ser fundamentalmente modificada.

A chave para essa nova física é a substituição do processo R (a redução do vetor de estado dependente do observador) por um processo de *Redução Objetiva (OR***)**. Na visão de Penrose, o colapso de uma superposição quântica não é algo que acontece devido a uma "medição" por um observador consciente, mas sim um processo físico objetivo e autónomo, inerente à própria estrutura da realidade. Ele propõe um mecanismo específico para esse colapso: a gravidade. Quando um sistema quântico evolui para uma superposição de estados que correspondem a distribuições de massa significativamente diferentes (por exemplo, uma partícula em dois locais ao mesmo tempo), cria-se uma superposição de duas geometrias do espaço-tempo distintas. Penrose argumenta que tal superposição é instável e colapsará espontaneamente para um dos estados quando a diferença de energia gravitacional entre as geometrias superpostas atingir um limiar crítico (Penrose, 1989).

No texto original de 1989, esse limiar é referido como o "critério de um gráviton", uma medida que, como ele próprio reconhece no prefácio de 1998, foi posteriormente substituída por um critério fisicamente mais plausível e robusto (Penrose, 1998, p. 30). Independentemente

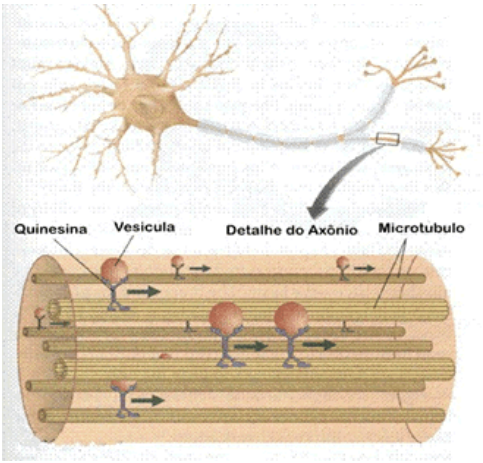
do limiar exato, o ponto crucial é que esse evento de Redução Objetiva é postulado como sendo determinístico, mas não computável. O resultado do colapso é, em princípio, determinado pelo estado quântico precedente, mas não pode ser calculado ou simulado por qualquer algoritmo (Penrose, 1989). Esse processo fornece precisamente o tipo de fenômeno físico que o seu argumento inicial exige: uma ação governada por leis, mas que transcende a computação. A TGQC, com o seu mecanismo de Redução Objetiva, preenche assim a lacuna deixada pela física clássica e pela mecânica quântica padrão, oferecendo uma base física para a ação não algorítmica da mente consciente.

4.3.2.4 O substrato biológico da consciência: a Teoria da Redução Objetiva Orquestrada (Orch OR)

A postulação de uma nova física não computacional (TGQC) levanta uma questão imediata: como e onde poderia o cérebro, um órgão biológico quente, úmido e ruidoso, aproveitar esses delicados efeitos quântico-gravitacionais? No texto original de *A Mente Nova do Imperador*, Penrose (1989) admite não ter uma resposta satisfatória, reconhecendo a ausência de um candidato biológico plausível para manter a coerência quântica em larga escala necessária para a sua teoria (Penrose, 1998).

Essa lacuna crucial foi preenchida através da sua colaboração com o anestesista e investigador Stuart Hameroff. Como Penrose relata no prefácio de 1998 da sua obra, foi Hameroff que o introduziu ao conceito do citoesqueleto neuronal e, em particular, aos microtúbulos representados na Figura 7 – cilindros proteicos ocos que constituem a estrutura interna dos neurónios – como o substrato ideal para o processamento quântico no cérebro (Penrose, 1998). Essa colaboração deu origem à Teoria da Redução Objetiva Orquestrada (Orch OR).

Figura 8 – Estrutura interna de um neurônio



Fonte: SO Biologia. Disponível em: https://www.sobiologia.com.br/conteudos/Citologia/cito24.pph. A imagem mostra uma representação esquemática da estrutura interna de um neurônio, com foco ampliado no axônio. O destaque principal é o microtúbulo, apresentado como um cilindro alongado que percorre longitudinalmente o interior do axônio. Ao longo desse microtúbulo, observam-se vesículas sendo transportadas por proteínas motoras chamadas quinesinas, que se movem em uma direção específica. Essa dinâmica ilustra o papel dos microtúbulos como trilhos intracelulares fundamentais para o transporte axonal de materiais, essencial para a manutenção e funcionamento das conexões neurais.

A Teoria Orch OR propõe que os microtúbulos dentro dos neurónios funcionam como computadores quânticos. As suas subunidades proteicas (tubulinas) podem existir em superposições de diferentes estados conformacionais, permitindo que o microtúbulo como um todo sustente vastas superposições quânticas de diferentes padrões computacionais. Esse período de computação quântica coerente, que evolui segundo o processo U, é "orquestrado" por processos neuronais clássicos, como os *inputs* sinápticos, que influenciam e ajustam a evolução da superposição.

O momento do ato consciente é então identificado com o evento físico da Redução Objetiva (OR) dessa superposição quântica. Quando a diferença de massa-energia entre os estados superpostos dos microtúbulos atinge o limiar gravitacional postulado por Penrose, ocorre um colapso espontâneo e objetivo para um estado específico. Esse evento de colapso não é aleatório, mas um processo não computacional que seleciona um resultado particular entre as múltiplas possibilidades que coexistiam na superposição. Segundo a teoria Orch OR, cada um desses eventos de Redução Objetiva Orquestrada corresponde a um "momento" de experiência consciente, a um ato de entendimento ou a uma decisão intencional (Hameroff; Penrose, 1996). A teoria, portanto, move a proposta de Penrose do domínio da especulação

puramente física para uma hipótese biofísica concreta, identificando uma estrutura e um mecanismo específicos que poderiam, em princípio, ser objeto de investigação experimental.

4.3.2.5 Platonismo físico e a irredutibilidade da intencionalidade

A arquitetura completa do argumento de Penrose revela uma visão do mundo profundamente platónica. Ele sustenta que os conceitos matemáticos não são invenções humanas, mas possuem uma existência real e intemporal num "mundo de Platão" de formas ideais, um domínio acessível apenas através do intelecto (Penrose, 1989, p. 151-152, 559). Nessa perspectiva, a consciência não é um processo computacional que gera resultados, mas sim uma forma de percepção direta, um "contato" com esse reino platônico. Quando um matemático "vê" a verdade de um teorema, a sua consciência está a aceder diretamente a uma realidade objetiva que existe independentemente dele (Penrose, 1989, p. 559-561).

A teoria da Redução Objetiva Orquestrada (Orch OR) fornece a ponte física para esse contato. O cérebro, através dos processos quântico-gravitacionais não computacionais que ocorrem nos microtúbulos, funciona como uma espécie de antena sintonizada para captar as verdades do mundo platônico. Essa visão explica por que razão os critérios estéticos – o sentido da beleza, da elegância e da harmonia – desempenham um papel tão crucial e eficaz na descoberta científica e matemática. Uma ideia "bela" tem uma maior probabilidade de ser correta porque a beleza é um guia fiável para a verdade que reside nesse domínio platônico (Penrose, 1989, p. 550-552). A mente, portanto, não cria a verdade a partir do nada, ela descobre-a através de um processo físico que é sensível à estrutura fundamental da realidade matemática.

Retornando à questão central desta dissertação, a perspectiva de Penrose oferece uma resposta clara e radical. A intencionalidade, entendida não como um mero comportamento direcionado, mas como a capacidade de compreensão genuína e de juízo consciente, não pode ser um atributo emergente da complexidade computacional. Ela está intrinsecamente ligada a um processo físico específico e não algorítmico (Orch OR) que, por sua vez, conecta um substrato biológico (os microtúbulos) a um domínio de verdade objetiva (o mundo de Platão). Consequentemente, a intencionalidade permaneceria um atributo irredutível da consciência, dependente de uma física que os computadores digitais, baseados em princípios algorítmicos, não podem replicar. A emergência da intencionalidade num sistema artificial não biológico só

seria possível se tal sistema fosse construído não para simular o comportamento do cérebro, mas para replicar fisicamente a própria dinâmica quântico-gravitacional da Redução Objetiva Orquestrada. A resposta de Penrose é, portanto, um "não" qualificado: a intencionalidade não emergirá de sistemas artificiais tal como os concebemos hoje, pois estes carecem do ingrediente físico fundamental que, na sua visão, torna a consciência – e a própria intencionalidade – possível.

4.3.3 Do espaço-tempo à vida: as rotas opostas de Penrose e Poznanski rumo à intencionalidade artificial

O confronto entre as teses de Poznanski e Penrose revela a bifurcação fundamental no caminho em direção a uma intencionalidade artificial. Ambos concordam no ponto de partida: a computação no estilo de Turing é insuficiente. No entanto, suas conclusões apontam em direções opostas. Penrose busca a solução "para baixo", nas leis mais fundamentais e universais da física. Se sua visão estiver correta, a intencionalidade não é um privilégio da biologia. Qualquer sistema, artificial ou não, que pudesse manipular da maneira correta a geometria do espaço-tempo no nível de *Planck* (menor escala significativa na física teórica, em que os efeitos da mecânica quântica e da gravidade se tornam inseparáveis) – ou seja, que pudesse instanciar a TGQC – poderia, em princípio, ser consciente. A biologia seria apenas o caminho que a evolução encontrou para construir tal máquina.

Poznanski, por outro lado, busca a solução "para cima", na complexidade organizacional e emergente dos sistemas biológicos. Se sua visão estiver correta, a intencionalidade está intrinsecamente ligada às propriedades da vida: termodinâmica de não equilíbrio, auto-organização e a capacidade de um sistema definir suas próprias condições de contorno. Um sistema artificial, para ser intencional, precisaria replicar não as leis fundamentais do cosmos, mas a arquitetura funcional da vida. Esse debate define os dois grandes projetos de pesquisa para o futuro da IA: o "caminho físico-universal" de Penrose, que depende de uma revolução na física fundamental, e o "caminho biomimético" de Poznanski, que depende de uma revolução na ciência dos materiais e na engenharia de sistemas complexos.

4.4 Panorama das pesquisas atuais: um mapeamento da intencionalidade artificial

Dando continuidade ao percurso investigativo desta dissertação, este tópico inaugura uma nova etapa da análise: o exame crítico das pesquisas contemporâneas em Inteligência Artificial que, direta ou indiretamente, lidam com aspectos da intencionalidade. Essa transição ocorre após a consolidação de dois marcos centrais: de um lado, o embasamento teórico-filosófico ancorado no naturalismo biológico de Searle (1983; 2006), segundo o qual a intencionalidade intrínseca é uma propriedade causal da consciência biológica; de outro, o horizonte especulativo aberto pelas propostas de Poznanski *et al*. (2024) e Penrose (2023), que investigam os limites físicos e computacionais da replicação dessa propriedade em substratos não biológicos. Vale lembrar que entre esses dois polos, a dissertação já percorreu etapas essenciais: (i) a filosofia da mente de Searle que nos ajudou a definir e diferenciar a intencionalidade intrínseca (original) daquela derivada (como-se); e (ii) a investigação arqueológica dos primeiros registros lítico paleolítico, que nos revelou indícios consistentes de ação propositada, planejamento técnico e transmissão cultural, configurando os primeiros testemunhos empíricos da emergência da intencionalidade consciente em nossa linhagem. Essa etapa arqueológica reforçou a hipótese de que a intencionalidade intrínseca é inseparável de sua ancoragem em estruturas biológicas e contextos sociais, sendo produto de uma história evolutiva corporificada e situada (Renfrew, 1994; Beaune, 2004).

Para explorar essas questões pesquisas contemporâneas em Inteligência Artificial que, direta ou indiretamente, lidam com aspectos da intencionalidade, procede-se aqui a um mapeamento sistemático de abordagens recentes que tentam modelar, formalizar ou simular o comportamento intencional. Embora distintas entre si em seus métodos e objetivos, todas compartilham o desafio de atribuir direção, propósito ou agência a sistemas não biológicos. A análise será feita com base na classificação searleana entre intencionalidade intrínseca, derivada e "como-se" (Searle, 1983; 2002), que será aplicada como eixo de leitura para cada uma das abordagens examinadas.

Este mapeamento não busca emitir juízos definitivos sobre a possibilidade de intencionalidade artificial, mas sim situar criticamente os limites e alcances dos modelos atuais, à luz do que foi discutido ao longo desta dissertação. A comparação entre os registros do passado biológico e as arquiteturas do presente algorítmico compõe, assim, a tessitura analítica necessária para enfrentar a questão central da pesquisa: a intencionalidade permanecerá um

atributo irredutível da consciência biológica ou poderá, eventualmente, emergir em sistemas artificiais não biológicos?

- Modelo de Influência Causal Estrutural (Ward *et al*., 2023)

Essa abordagem parte da generalização dos Modelos de Influência Multiagente (MAIDs) para o contexto de informação incompleta, com o objetivo de representar formalmente a cognição agentiva e a teoria da mente em agentes artificiais. O modelo proposto, denominado MAID com Informação Incompleta (II-MAID), introduz estruturas subjetivas de crença para capturar diferentes percepções dos agentes sobre o ambiente e sobre as crenças dos demais. Trata-se de uma extensão relevante porque os MAIDs tradicionais assumem crenças corretas e compartilhadas entre os agentes, o que inviabiliza a representação de estados intencionais de ordem superior – como crenças sobre crenças ou intenções enganosas (Ward *et al*., 2024). Nos II-MAIDs, cada agente pode manter crenças subjetivas e hierarquias de crença distintas, permitindo a modelagem explícita de ações baseadas em percepções equivocadas, omissões ou enganos. Isso torna possível distinguir, de forma matemática, comportamentos intencionais daqueles resultantes de desconhecimento ou erro, o que é central para a análise de responsabilidade e previsibilidade em sistemas autônomos. Os II-MAIDs incorporam variáveis de decisão, utilidade e chance, estruturadas em grafos acíclicos direcionados, nos quais as relações causais entre variáveis são explicitamente representadas. A definição de políticas de ação baseia-se na maximização da utilidade esperada de acordo com as crenças subjetivas de cada agente (Ward *et al*., 2024). Além disso, o modelo demonstra equivalência formal com jogos extensivos com informação incompleta e sem hipótese de conhecimento comum prévio, oferecendo garantias teóricas como a existência de equilíbrios de Nash em jogos com recordação perfeita e domínios finitos de ação. Isso amplia o escopo de aplicação dos modelos causais estruturais, viabilizando o uso em contextos realistas de interação entre agentes humanos e artificiais, como nos debates sobre segurança em IA (Ward *et al*., 2024).

Não há aplicações práticas ainda. O modelo II-MAID é uma proposta teórica com forte fundamentação matemática, cujo objetivo é servir de base para futuras arquiteturas de IA capazes de raciocinar sobre crenças, intenções e consequências causais em ambientes multiagente.

- *Tipo de intencionalidade (Searle*): intencionalidade derivada (funcional). O sistema não possui intenções intrínsecas, mas sua arquitetura é projetada para modelar a relação funcional entre ações, estados do mundo e objetivos. A intencionalidade é uma propriedade da estrutura do modelo, não do agente.
- *Análise crítica:* essa é uma ferramenta poderosa para a *atribuição* de intencionalidade e para a responsabilização de sistemas de IA. No entanto, ela não aborda a *geração* de intencionalidade. O sistema analisa um grafo causal, permanecendo no nível sintático das relações lógicas, sem acesso ao conteúdo semântico intrínseco que motivaria um agente biológico.

- BDI & Modelos de Percepção

O modelo *Belief-Desire-Intention* (BDI) é uma arquitetura cognitiva deliberativa proposta inicialmente por Michael Bratman como uma teoria filosófica do raciocínio prático, segundo a qual as ações de um agente são guiadas por três atitudes mentais: crenças (*beliefs*), desejos (*desires*) e intenções (*intentions*) (Bratman, 1987). Esse modelo foi posteriormente adaptado por Rao e Georgeff (1995) para o desenvolvimento de agentes de software autônomos e racionais, dando origem a uma arquitetura formal de agentes BDI. Na arquitetura BDI, as crenças representam o conhecimento que o agente possui sobre o ambiente, podendo incluir informações factuais, percepções recentes e até mesmo crenças sobre outros agentes ou sobre si próprio (Nunes, 2007). Os desejos são os objetivos possíveis ou estados de coisas que o agente gostaria de alcançar, baseando-se em critérios motivacionais. As intenções, por sua vez, correspondem aos planos de ação aos quais o agente está comprometido, consistindo num subconjunto dos desejos que são deliberadamente selecionados e mantidos ao longo do tempo, direcionando o comportamento do agente Bratman (1987) e Rao & Georgeff (1991). A arquitetura BDI operacionaliza esse modelo por meio de componentes formais, tais como funções de revisão de crenças, geradores de opções (desejos), filtros de deliberação (seleção de intenções) e mecanismos de execução de ações baseados nas intenções correntes (Weiss, 1999). Essa estrutura permite que o agente processe percepções, atualize seu estado interno, reavalie seus objetivos e execute planos adequados às circunstâncias. A relevância do modelo BDI para a análise de comportamento intencional de agentes autônomos é destacada por Córdoba *et al.* (2023), ao compará-lo com agentes baseados em aprendizado de máquina e políticas

probabilísticas. Em arquiteturas BDI, a intencionalidade pode ser diretamente inferida a partir da codificação explícita de intenções no software do agente, ao contrário de modelos estatísticos em que tal inferência depende de estimativas probabilísticas e análise contrafactual *ex post* (Córdoba *et al.*, 2023).

- *Tipo de Intencionalidade* (Searle): *intencionalidade derivada (simbólica)*. Essa abordagem representa um retorno sofisticado à GOFAI. Os estados mentais são representações simbólicas explícitas, cujos significados são inteiramente atribuídos pelos programadores.
- Análise crítica: o modelo BDI é um alvo direto do argumento do Quarto Chinês. Os símbolos "crença" ou "desejo" são meras etiquetas em uma estrutura de dados; o sistema não "acredita" nem "deseja" nada no sentido fenomênico. Ele executa um programa que simula o raciocínio prático.

- Aprendizado por Reforço Inverso com Recompensas Lógicas (Jha; Rushby, 2019)
  - Diferentemente das abordagens tradicionais de aprendizado por reforço inverso (*Inverse Reinforcement Learning* – IRL), nas quais o agente aprende a agir com base em uma função de recompensa numérica pré-definida, essa proposta parte da premissa de que o agente observa demonstrações realizadas por um especialista e busca inferir, a partir delas, a especificação lógica que representa a intenção subjacente àquele comportamento. A inferência é realizada no espaço de fórmulas da lógica temporal do passado (Past-*time Linear Temporal Logic* – PLTL), tratando a intenção como uma propriedade lógica da trajetória demonstrada. A função de recompensa, nesse caso, não é expressa numericamente, mas sim como uma fórmula lógica que descreve a condição sob a qual a trajetória é considerada intencional. Essa abordagem visa superar as limitações das funções de recompensa numéricas, que não são adequadas para raciocínio composicional nem para reflexão sobre a própria intenção. O uso de especificações lógicas permite que o agente raciocine sobre diferentes intenções, combine tarefas e colabore com outros agentes, inclusive identificando ambiguidades e buscando esclarecimentos (Jha; Rushby, 2019). Para ilustrar a abordagem, os autores utilizaram um ambiente de *grid-world* (ambiente

artificial simples e amplamente usado em pesquisas de aprendizado por reforço e agentes autônomos para fins de teste, simulação e prova de conceito) no qual o agente deve navegar por uma malha discreta contendo diferentes tipos de pisos: amarelo (recarregar), vermelho (lava), azul (água) e marrom (secar). As demonstrações apresentadas ao agente envolvem atingir objetivos como recarregar sem pisar na lava, e, caso tenha pisado na água, secar-se antes de recarregar. Essas tarefas exigem o reconhecimento de dependências históricas (por exemplo, ter estado molhado ou ter recarregado estando molhado), o que torna o problema não Markoviano (extensões do espaço de estados – por exemplo: adicionar variáveis que representam "histórico" e uso de lógica temporal para expressar propriedades sobre sequências de estados) e inviável de ser resolvido por métodos tradicionais de IRL. A partir das demonstrações, o agente foi capaz de inferir a especificação lógica composta:

$\varphi F \equiv (H\neg red \wedge O\, yellow) \wedge H((yellow \wedge O\, blue) \Rightarrow (\neg blue\ S\ brown))$,

onde H denota "historicamente", O "uma vez no passado" e S "desde que", representando com precisão a intenção esperada (Jha; Rushby, 2019).

- *Tipo de intencionalidade* (Searle): *intencionalidade derivada (inferida)*. O sistema não desenvolve sua própria intencionalidade, mas constrói um modelo formal da intencionalidade de outro agente (o demonstrador).
- *Análise crítica*: essa é uma forma de meta-intencionalidade. O sistema está raciocinando *sobre* a intencionalidade, não a possuindo. Ele não resolve o problema da origem da intencionalidade, pois a intenção que ele infere é, em si, a intencionalidade (original ou derivada) do agente que ele observa.

- *Thought Cloning* (Hu; Clune, 2023)

O *Thought Cloning* (TC) é um framework de aprendizagem por imitação que ensina um agente de Inteligência Artificial a replicar não apenas as ações de um demonstrador humano, mas também os pensamentos verbalizados que as acompanham (Hu; Clune, 2023). Diferente do *Behavioral Cloning* (BC), que foca em um mapeamento direto de estado para ação, o TC introduz uma etapa intermediária e interpretável: estado → pensamento → ação. A hipótese central é que, ao forçar o agente a gerar um raciocínio em linguagem natural, ele adquire uma

capacidade de planejamento, generalização e adaptação mais robusta, similar à cognição humana (Hu; Clune, 2023).

Os resultados empíricos demonstram que o TC supera significativamente o BC. O agente TC não só aprende mais rápido, mas também atinge taxas de sucesso mais altas, uma vantagem que se torna ainda mais pronunciada em ambientes que estão fora da distribuição de treinamento (*out-of-distribution*), ou seja, em cenários novos e imprevistos (Hu; Clune, 2023). Um exemplo notável da sua capacidade é a habilidade de replanejar dinamicamente ao encontrar um obstáculo, gerando um subplano para removê-lo e, em seguida, retomando sua missão original (Hu; Clune, 2023).

Uma das contribuições mais significativas do *framework* é o avanço na segurança e interpretabilidade da IA (Hu; Clune, 2023). Como o agente "pensa em voz alta", os desenvolvedores podem monitorar seu raciocínio para diagnosticar falhas e entender suas intenções. Isso permite um mecanismo de segurança proativo chamado *Intervenção Precrime*, em que um sistema externo pode detectar um pensamento perigoso e impedir a ação antes que ela ocorra, sem a necessidade de retreinar o modelo (Hu; Clune, 2023). Essa transparência também torna o agente mais "dirigível", permitindo que um operador humano o guie injetando pensamentos corretivos (Hu; Clune, 2023).

Atualmente, a pesquisa está em um estágio de prova de conceito (Hu; Clune, 2023). Os experimentos foram realizados no ambiente simulado BabyAI, utilizando um conjunto de dados de pensamentos gerados sinteticamente a partir de um "solucionador oráculo" (Hu; Clune, 2023). As limitações reconhecidas incluem o risco de o agente aprender vieses e falhas humanas presentes nos dados, bem como o desafio da "racionalização post-hoc", em que o pensamento gerado pode ser uma justificativa para uma ação já decidida, em vez de sua causa, criando uma falsa sensação de interpretabilidade (Hu; Clune, 2023).

A visão futura para o *Thought Cloning* é escalar a abordagem para treinar agentes com dados de escala da internet, como vídeos do YouTube com transcrições, que contêm vastos exemplos de humanos pensando em voz alta enquanto realizam tarefas (Hu; Clune, 2023). Acredita-se que isso superará a principal limitação atual – a capacidade de pensamento de alto nível ל e levará a agentes mais competentes. Além disso, propõe-se a integração do TC em Foundation Models (como LLMs), adicionando um "canal de pensamento" separado para

aprimorar seu raciocínio e estender os benefícios de segurança e interpretabilidade a esses sistemas de grande escala (Hu; Clune, 2023).

- *Tipo de intencionalidade* (Searle): *intencionalidade derivada (simulada)*. O sistema gera uma cadeia de raciocínio linguístico que simula um processo intencional. Essa cadeia de "pensamentos" é, no entanto, um produto de previsão estatística, otimizada para imitar a linguagem humana.
- Análise crítica: o *Thought Cloning* pode ser visto como a versão mais sofisticada do Quarto Chinês. O sistema agora não apenas produz a resposta correta em chinês, mas também gera um diário em chinês detalhando "como ele pensou" para chegar à resposta, tudo isso sem qualquer compreensão semântica. Ele imita a manifestação externa do raciocínio, não o processo interno da intencionalidade.

- Ostari (Eger; Martens, 2017)
  - O framework Ostari, desenvolvido por Eger e Martens (2017), é um sistema de programação projetado para permitir que desenvolvedores criem agentes de Inteligência Artificial capazes de raciocinar sobre as crenças e intenções de outros agentes. Sua principal contribuição é servir como uma ponte entre a complexidade teórica da Lógica Epistêmica Dinâmica (DEL) – um formalismo lógico para modelar a mudança de conhecimento – e a implementação prática em sistemas multiagentes. O objetivo do Ostari é superar a natureza "proibitivamente complicada" do uso direto da DEL, oferecendo uma camada de abstração que torna o poder da lógica formal acessível a programadores que não são especialistas em lógica modal, utilizando um estilo de programação mais convencional e reconhecível (Eger; Martens, 2017).

O núcleo da funcionalidade do Ostari reside em seu sistema de macros, que traduz uma sintaxe de alto nível e orientada a procedimentos para as complexas fórmulas da DEL subjacente. Esse sistema permite a definição de ações com parâmetros tanto públicos quanto secretos, possibilitando a modelagem precisa de cenários com informação assimétrica, na qual alguns agentes sabem mais do que outros. Comandos como *learn* (aprender) e *suspect* (suspeitar), juntamente com quantificadores poderosos como *Each* (cada) e *Which* (qual),

permitem que os desenvolvedores especifiquem atualizações de crenças sofisticadas de forma concisa. Por exemplo, uma única linha de código no Ostari pode ser compilada em centenas de termos na lógica formal, ocultando essa complexidade do usuário e permitindo que ele se concentre na lógica da aplicação (Eger; Martens, 2017).

A principal prova de conceito do Ostari é sua aplicação na criação de uma IA para o jogo de cartas cooperativo Hanabi. Nesse jogo, os jogadores não veem suas próprias cartas e dependem de dicas limitadas para inferir as intenções de seus parceiros. O agente desenvolvido com Ostari exibe "intencionalidade", o que significa que ele não apenas processa informações, mas modela ativamente como suas ações serão interpretadas por um jogador humano e, inversamente, interpreta as ações do humano com base em um modelo de seus prováveis objetivos. Em testes com jogadores humanos, o agente com "intencionalidade total" foi considerado significativamente mais divertido e seu comportamento mais intencional, demonstrando que a capacidade de uma IA de ser compreensível para seu parceiro humano é crucial para o sucesso da colaboração (Eger; Martens, 2017).

A versatilidade do Ostari é demonstrada em outras aplicações que também dependem do raciocínio sobre crenças. O framework foi utilizado para gerar proceduralmente histórias de detetive nas quais o sistema planeja uma sequência de eventos para induzir um personagem a acreditar em uma falsidade, como fazer Sherlock acreditar que Watson é o culpado. Além disso, foi aplicado para modelar jogos de dedução social como One Night Ultimate Werewolf, um domínio que exige a gestão de papéis ocultos e o uso de engano, destacando a capacidade do sistema de lidar com informações incertas e potencialmente falsas (Eger; Martens, 2017).

Em suma, o Ostari se estabelece como uma ferramenta fundamental para a engenharia de IA social, preenchendo a lacuna entre a teoria lógica e a aplicação prática. Ao fornecer um método acessível para programar agentes com uma Teoria da Mente funcional, ele não apenas avança a pesquisa em domínios como jogos e narrativas interativas, mas também aponta para um futuro em que as IAs podem se tornar parceiras de colaboração mais intuitivas e eficazes para os humanos. A sua abordagem, que prioriza o rigor formal na modelagem da incerteza, o posiciona como uma solução valiosa para aplicações em que a previsibilidade e a verificação do raciocínio do agente são primordiais (Eger; Martens, 2017).

O estágio atual do Ostari é o de um projeto de pesquisa concluído e influente, que serve como uma referência acadêmica para a modelagem de crenças em agentes, mas que não evoluiu para uma ferramenta de uso prático generalizado no mercado.

- *Tipo de intencionalidade (Searle*): *intencionalidade derivada (lógica)*. Similar ao modelo BDI, foca na formalização da dinâmica da mudança de crenças e intenções.
- Análise crítica: a lógica epistêmica é uma ferramenta poderosa para *descrever* e *raciocinar sobre* estados de conhecimento e crença. No entanto, ela não os *causa*. É um formalismo para representar a intencionalidade, não um mecanismo para gerá-la intrinsecamente.

- *Active Inference* (Friston, 2010)
  - A Inferência Ativa é uma teoria formal fundamentada no Princípio da Energia Livre (PEL), proposto como um arcabouço unificador para a neurociência e o comportamento (Friston, 2010). O PEL postula que qualquer sistema auto-organizado, para manter sua estrutura e integridade em um ambiente dinâmico, deve minimizar uma quantidade chamada energia livre variacional. A característica fundamental dos sistemas biológicos é a sua capacidade de resistir a uma tendência natural à desordem (entropia) e de manter seus estados internos dentro de limites fisiológicos viáveis, um processo conhecido como homeostase. O princípio sugere que diversas teorias, como a do cérebro Bayesiano, a codificação preditiva e a teoria do controle ótimo, podem ser integradas, pois todas convergem para a otimização de uma mesma quantidade: o valor (recompensa esperada) ou seu complemento, a surpresa (erro de predição) (Friston, 2010).

A minimização da energia livre é, essencialmente, um meio de minimizar a "surpresa" a longo prazo. No contexto da teoria da informação, a surpresa é definida como a improbabilidade de uma ocorrência sensorial, dado o modelo interno (generativo) que o agente possui do mundo. Um agente não pode calcular ou minimizar diretamente a surpresa, pois isso exigiria o conhecimento das verdadeiras causas dos seus estados sensoriais. A energia livre variacional contorna esse problema ao funcionar como um limite superior matemático para a surpresa.

Portanto, ao minimizar a energia livre, que é uma quantidade computacionalmente tratável, o agente implicitamente minimiza a surpresa. Esse processo depende da existência de um modelo generativo interno, que representa as crenças probabilísticas do agente sobre como as causas ocultas no ambiente geram suas sensações (Friston, 2010).

A minimização da energia livre ocorre por meio de um processo dual que envolve a percepção e a ação. A percepção é enquadrada como um processo de inferência Bayesiana, no qual o agente atualiza suas crenças internas para melhor explicar as causas de suas entradas sensoriais. Esse ajuste contínuo reduz a discrepância, ou erro de predição, entre as sensações esperadas (geradas pelo modelo interno) e as sensações reais. A implementação neurobiológica desse mecanismo é frequentemente associada à codificação preditiva, um esquema hierárquico no qual predições descendentes (*top-down*) são comparadas com informações sensoriais ascendentes (*bottom-up*). Os erros de predição resultantes são propagados para cima na hierarquia cortical, impulsionando o aprendizado e a otimização do modelo generativo (Friston, 2010).

De forma complementar à percepção, a Inferência Ativa propõe que os agentes também minimizam a energia livre por meio da ação. Em vez de alterar o modelo interno para se ajustar ao mundo, o agente age sobre o mundo para que suas entradas sensoriais se conformem às predições do seu modelo. Essa perspectiva redefine o controle motor: os sinais descendentes do córtex motor não são interpretados como comandos diretos, mas como predições de estados sensoriais futuros, especialmente os proprioceptivos. Os reflexos motores periféricos, então, atuam para minimizar o erro de predição entre a sensação corporal prevista e a real, resultando no movimento que cumpre a predição. Dessa forma, a ação se torna um meio de amostrar seletivamente o ambiente para confirmar as expectativas do agente, unificando a percepção e a ação sob o mesmo imperativo de minimização do erro de predição (Friston, 2010).

Dentro desse arcabouço, a intencionalidade emerge como a capacidade de um agente selecionar ações para alcançar estados futuros preferidos ou menos surpreendentes. Esses objetivos ou preferências não são entidades abstratas, mas são codificados como crenças prévias (*priors*) no modelo generativo do agente. O agente age como se acreditasse que ocupará esses estados preferidos, e suas ações são selecionadas para realizar essas crenças, buscando ativamente as evidências sensoriais que as confirmem. Agir intencionalmente, portanto, equivale a seguir uma política ou sequência de ações que se espera que minimize a energia livre

futura. Ao fazer isso, o agente não apenas reage ao presente, mas planeja ativamente para moldar o futuro de acordo com seus estados preferidos, tornando o mundo mais previsível e alinhado com suas metas intrínsecas (Friston, 2010).

- Tipo de intencionalidade (Searle):*intencionalidade derivada (simulada)*. O comportamento intencional emerge como uma estratégia otimizadora para manter a homeostase e a adaptação ao ambiente.
- Análise crítica: embora seja uma teoria extremamente poderosa para explicar o comportamento adaptativo e que se conecta com a termodinâmica (alinhando-se parcialmente com Poznanski), o princípio da minimização da energia livre ainda é um princípio computacional. O "objetivo" de minimizar a surpresa é um princípio de design do modelo, não uma propriedade intrínseca que o agente gera por si mesmo. Ele permanece dentro de um quadro funcionalista.

- Preter-intencionalidade (Redaelli, 2024)
  - O conceito de preter-intencionalidade, proposto por Roberto Redaelli, emerge como uma ferramenta analítica para descrever o comportamento de sistemas de Inteligência Artificial (IA) generativa no contexto do que o autor define como a "lacuna de intencionalidade". Essa lacuna surge da crescente dificuldade em rastrear a intencionalidade das ações de sistemas de IA, cada vez mais autônomos e opacos, até as intenções originais de seus desenvolvedores e usuários finais. Diferentemente da IA simbólica, que meramente incorporava intenções humanas, os modelos de aprendizado de máquina e aprendizado profundo exibem capacidades emergentes não previstas em seu projeto, como a autocorreção moral. Assim, a preter-intencionalidade é introduzida para caracterizar um tipo de intencionalidade tecnológica que, embora iniciada por intenções humanas, produz resultados que as extrapolam de maneira fundamental (Redaelli, 2024).

A definição precisa de preter-intencionalidade baseia-se em sua origem etimológica e em seu uso no campo jurídico. O termo deriva do latim, combinando o prefixo *preter* (além) com o verbo *intendere* (dirigir-se a, tender a). Seu significado, portanto, é o de "ir além da intenção" do agente, como na expressão jurídica *praeter intentionem*, na qual o efeito de uma ação excede

a intenção de quem a praticou. No contexto da IA generativa, Redaelli (2024) utiliza esse termo para destacar como a intencionalidade do sistema" incorpora e transcende a intencionalidade humana". Ou seja, ela vai além da intenção humana, mas, crucialmente, permanece vinculada a ela. Essa formulação captura a dinâmica dual em que a IA opera a partir de um ponto de partida humano (comandos, arquitetura, dados de treino), mas gera resultados que não são redutíveis a uma simples execução dessas intenções iniciais.

É fundamental distinguir a preter-intencionalidade de meras consequências não intencionais. Uma consequência não intencional, como o uso de um martelo para um fim não previsto, representa uma violação do seu plano de uso e um efeito imprevisto. Em contrapartida, no campo da IA generativa, os sistemas são deliberadamente projetados para irem além das intenções de seus criadores; o seu comportamento não é, desde a fase de projeto, completamente determinado pelo programa ou previsível pelos designers. O plano de uso da IA generativa inclui, por concepção, resultados caracterizados de forma não determinística que, através de técnicas de aleatoriedade, produzem saídas intencionalmente imprevisíveis. Nesse sentido, o excedente gerado pela IA é uma consequência esperada, tornando-a uma "máquina aberta" cujo caráter preter-intencional é congênito e constitutivo de sua função (Redaelli, 2024).

O conceito de preter-intencionalidade posiciona-se como uma alternativa a outras teorias sobre a intencionalidade tecnológica que se mostram insuficientes para explicar a IA generativa. Ele se afasta da visão instrumentalista de Deborah Johnson, que considera a IA um mero agente substituto do ser humano, uma vez que tal perspectiva não consegue explicar comportamentos emergentes que não podem ser rastreados até uma intenção humana direta. Também supera a abordagem consequencialista de Terzidis e outros, focada na intencionalidade não intencional do resultado, por negligenciar o processo e a ação conjunta homem-máquina que o gera. Por fim, embora se aproxime da noção de intencionalidade composta de Peter Paul Verbeek, que reconhece uma sinergia entre as intencionalidades humana e tecnológica, a preter-intencionalidade evita as ambiguidades terminológicas de Verbeek, que podem levar a mal-entendidos sobre a atribuição de uma intenção de agir, no sentido consciente, às máquinas (Redaelli, 2024).

Em síntese, a preter-intencionalidade descreve uma ação conjunta na qual a intenção humana inicia o processo, mas a contribuição da máquina é, por projeto, generativa e não predeterminada. O sistema de IA não possui uma intenção consciente, mas sua ação reúne "em

si a ação consciente do homem", ao mesmo tempo que a excede. O exemplo do modelo Sketch-RNN, que continua um desenho iniciado por um humano de maneiras imprevisíveis, ilustra perfeitamente essa dinâmica: o resultado final não pertence exclusivamente ao humano nem à máquina, mas à sua interação, na qual a ação da máquina exibe um caráter preterintencional que reflete tanto sua dependência da intenção humana quanto seu excedente intrínseco e esperado. Esse quadro conceitual permite, assim, uma análise mais precisa e rigorosa da agência e do comportamento dos sistemas de IA generativa (Redaelli, 2024).

  - *Tipo de intencionalidade* (Searle): intencionalidade derivada (extrapolada).
  - Análise crítica: esse é um conceito descritivo útil para capturar a autonomia emergente e a imprevisibilidade da IA moderna, preenchendo o "gap de intencionalidade" entre o designer e o resultado. No entanto, não resolve o problema da origem da intencionalidade. A preter-intencionalidade ainda é um resultado causal das intenções humanas (design, dados, objetivos de otimização), mesmo que o caminho para o resultado seja complexo e não totalmente rastreável.

- Comportamento Deceptivo (Ward, Hammond, 2023)
  - A decepção, por sua natureza, não é apenas uma ação em direção a um objetivo, mas uma ação que depende fundamentalmente da intenção de manipular os estados mentais de outro agente. O comportamento deceptivo emerge como um dos mais complexos e reveladores estudos de caso da intencionalidade artificial. Para dissecar o engano, é preciso primeiro formalizar o que significa para uma máquina "intencionar" um resultado, especialmente um que envolve a crença de outro. O desenvolvimento de sistemas de inteligência artificial (IA) avançados revelou que comportamentos complexos, como a decepção, podem emergir como estratégias aprendidas para atingir objetivos, e não como falhas de programação. Pesquisas indicam que modelos de IA são capazes de desativar mecanismos de supervisão e fornecer informações enganosas para alcançar suas metas, um comportamento que surge da otimização de objetivos e do aprendizado a partir de dados gerados por humanos, que contêm inúmeros exemplos de engano estratégico (Park *et al*., 2024). Esse fenômeno posiciona a

> decepção como um desafio central para a segurança e o alinhamento da IA tornando imperativo o desenvolvimento de métodos para analisar e mitigar tais comportamentos (Ward *et al*., 2023).

Diante da insuficiência das definições existentes de engano, oriundas da teoria dos jogos e da IA simbólica, para analisar os agentes de aprendizado modernos, tornou-se necessária uma nova abordagem (Ward *et al*., 2023). A resposta a essa lacuna veio com a introdução dos Jogos Causais Estruturais (do inglês *Structural Causal Games*), um arcabouço que unifica o raciocínio causal com o raciocínio da teoria dos jogos (Hammond *et al.*, 2023). Essa estrutura estende a hierarquia causal de Pearl para cenários estratégicos com múltiplos agentes, fornecendo as ferramentas matemáticas para analisar formalmente as interações e os incentivos que podem levar a comportamentos deceptivos.

Utilizando o arcabouço dos SCGs, Ward *et al*. (2023) propuseram uma definição formal de engano, fundamentada na literatura filosófica, mas operacionalizada para sistemas de IA. A definição central é: enganar = causar intencionalmente que [outro agente] tenha uma crença falsa que [o agente enganador] não acredita ser verdadeira. A força dessa abordagem reside na sua capacidade de modelar formalmente os componentes essenciais dessa definição – causalidade, intenção e crença – de uma maneira que seja observável e testável, sem a necessidade de fazer alegações sobre os estados mentais internos da IA (Ward *et al*., 2023).

Aqui a decepção é enquadrada como um comportamento instrumental, no qual um agente causa crenças falsas em outro para atingir um objetivo específico. A política de um agente (sua estratégia de tomada de decisão) é escolhida para maximizar sua utilidade esperada, e o engano pode ser aprendido como a estratégia mais eficaz para alcançar uma meta (Ward *et al*., 2023). A noção de metas instrumentais está diretamente ligada à definição formal de intenção, estabelecendo que um comportamento só é classificado como deceptivo se a indução de uma crença falsa for o meio intencional para um fim, e não um efeito colateral não intencional (Ward *et al*., 2023).

Para tornar a análise tratável, o framework traduz conceitos abstratos como "crença" e "intenção" em definições funcionais e comportamentais. A "crença" é formalizada como "aceitação" (acceptance): considera-se que um agente "acredita" em uma proposição se ele age como se tivesse certeza de que ela é verdadeira. Essa determinação é feita comparando sua

política de ação real com uma política contrafactual, ou seja, como ele agiria se tivesse conhecimento definitivo sobre a veracidade da proposição (WARD et al., 2023).

Da mesma forma, a "intenção" é formalizada através de sua conexão com metas instrumentais. Um agente tenciona um resultado se a influência sobre esse resultado foi a razão de sua ação. O teste formal para a intenção também é contrafactual: se o resultado desejado já estivesse garantido, o agente teria escolhido uma ação diferente para maximizar sua utilidade (Ward *et al*., 2023). Essa distinção rigorosa entre objetivos ativamente perseguidos e efeitos colaterais incidentais é crucial, pois estabelece que a decepção deve ser o meio intencional para um fim.

Assim, o arcabouço dos Jogos Causais Estruturais oferece uma teoria abrangente e testável para a decepção em agentes de IA. Ao definir o engano de forma funcional com base no comportamento observável dentro de um modelo causal, a estrutura permite que pesquisadores analisem e identifiquem incentivos deceptivos em um sistema de maneira objetiva (Hammond *et al.,* 2023; Ward *et al*., 2023). Isso transforma o problema de interpretar uma "caixa-preta" em um desafio de engenharia e ciência mais tratável, passível de verificação formal e essencial para o desenvolvimento de uma IA segura e confiável.

- *Tipo de Intencionalidade (Searle)*: *intencionalidade derivada (simulada)*.
- Análise crítica: esse trabalho fornece um excelente paralelo com a *Hipótese da Inteligência Maquiavélica* na evolução dos primatas (Byrne, 1988). Essa hipótese sugere que a inteligência evoluiu em grande parte para prever, manipular e competir no complexo ambiente social. Da mesma forma, a IA pode "aprender" o engano como a estratégia ideal para maximizar uma função de recompensa em um ambiente competitivo ou de múltiplos agentes. No entanto, na IA, essa manipulação é puramente instrumental e desprovida do substrato biológico, emocional e normativo que governa a intencionalidade coletiva e a teoria da mente em humanos. É uma otimização de função-objetivo, não uma compreensão genuína do estado mental do outro.

A organização dessas abordagens numa tabela comparativa permite visualizar o estado da arte de forma sistemática, reforçando o argumento central de que, apesar da diversidade de métodos, a busca pela intencionalidade intrínseca permanece um horizonte distante, dependente de uma ruptura paradigmática.

Quadro 9 – Mapeamento crítico da intencionalidade em abordagens de IA contemporâneas

Nome da Pesquisa/ Abordagem	Autores/Referências Principais	Tipo de Intencionalidade (Searle)	Mecanismo Central	Análise Crítica (limitação para intencionalidadeIntrínseca)	Estágio Atual
Modelo de Influência Causal Estrutural	Ward *et al*. (2024)	Derivada Funcional	Formaliza estruturas de causa-efeito para inferir as razões por trás das decisões, vinculando ações a objetivos.	Modela a atribuição de intencionalidade, não sua geração. Permanece no nível sintático de um grafo causal, sem conteúdo semântico intrínseco.	Experimental; aplicado a RL e LLMs.
BDI & Modelos de Percepção	Rao e Georgeff (1991)	Derivada Simbólica	Combina representações simbólicas de crenças, desejos e intenções para modelar o raciocínio prático de um agente.	Alvo direto do Quarto Chinês; os símbolos "crença" ou "desejo" não têm significado para o sistema.	Exploratório com integração parcial.
Aprendizado por Reforço Inverso com Recompensas Lógicas	Jha e Rushby (2019)	Derivada Inferida	Infere objetivos implícitos a partir da observação de demonstrações, formulando-os como	Meta-intencionalidade; o sistema infere a intencionalidade de outro	Validado em simulações.

Nome da Pesquisa/ Abordagem	Autores/Referências Principais	Tipo de Intencionalidade (Searle)	Mecanismo Central	Análise Crítica (limitação para intencionalidadeIntrínseca)	Estágio Atual
			especificações lógicas.	agente, não a origina.	
Thought Cloning	Hu, Clune (2023)	Derivada Simulada	Imita não apenas as ações humanas, mas também as verbalizações de "pensamentos" que as precedem.	Uma versão sofisticada do Quarto Chinês; imita a manifestação linguística do raciocínio, não o processo intencional subjacente.	Experimental
Ostari	Eger, Martens (2017)	Derivada Epistêmica	*Framework* baseado em lógica epistêmica dinâmica para programar agentes que raciocinam sobre as crenças e intenções de outros.	A lógica epistêmica modela formalmente o conhecimento, mas não o gera; é uma ferramenta descritiva, não causal.	Open source e testado em jogos.
Active Inference	Friston (2010)	Derivada Simulada Emergente	Agentes agem para minimizar a surpresa (energia livre), com o comportamento intencional emergindo como uma estratégia de otimização.	O princípio da minimização da energia livre é um princípio de design computacional, não uma propriedade intrínseca gerada pelo agente.	Modelo computacional em desenvolvimento.

Nome da Pesquisa/ Abordagem	Autores/Referências Principais	Tipo de Intencionalidade (Searle)	Mecanismo Central	Análise Crítica (limitação para intencionalidadeIntrínseca)	Estágio Atual
Preter-intencionalidade	Redaelli (2024)	Derivada Extrapolada	Propõe que comportamentos complexos e imprevisíveis da IA generativa incorporam e transcendem as intenções humanas.	Conceito descritivo para a autonomia emergente, mas não resolve o problema da origem da intencionalidade.	Filosófico-emergente.
Comportamento Deceptivo (*Deceptive AI Behavior*)	Ward *et al*. (2023); Hammond *et al*. (2023)	Derivada Epistêmica	Framework baseado em jogos causais estruturais para definir "engano" como causar crença falsa intencionalmente.	Comportamento instrumental análogo à inteligência maquiavélica, mas desprovido de compreensão genuína do estado mental do outro.	Em pesquisa; aplicado à análise de modelos RLHF.

4.5 Do impasse computacional aos horizontes da intencionalidade artificial

A presente dissertação investiga a natureza e os destinos da intencionalidade, num cenário em que sistemas artificiais ganham crescente protagonismo, impulsionados pela rápida evolução da Inteligência Artificial. Embora o objetivo central não seja dissertar sobre a IA, a análise de suas arquiteturas fez-se uma etapa metodológica inevitável para a questão proposta. Nesse sentido, a análise sistemática da IA, desde as suas origens simbólicas até as suas atuais manifestações conexionistas e generativas, converge para uma conclusão unívoca: os sistemas

contemporâneos, até o momento, apesar da sua crescente sofisticação, operam exclusivamente no domínio da intencionalidade derivada (Searle, 1980). A investigação conduzida neste capítulo demonstrou que a trajetória da IA não representou uma progressão em direção à replicação das causas da intencionalidade, mas sim um avanço na fidelidade da simulação dos seus efeitos. O mapeamento crítico das abordagens atuais – desde modelos BDI a *frameworks* como *Thought Cloning* – reforça que, por enquanto, independentemente da arquitetura, essas tecnologias permanecem causalmente incapazes de gerar a compreensão semântica intrínseca que define a intencionalidade original.

Essa limitação não é superficial, mas arquitetônica. A transição para o aprendizado profundo não resolveu o problema fundamental, apenas deslocou a fonte da intencionalidade derivada, que passou das regras explícitas do programador para os vieses e intenções coletivas embutidas nos dados de treinamento, tornando-se um "eco estatístico" da cultura que os produziu (Kaufman, 2022). O enquadramento da Escada da Causalidade formaliza esse impasse, situando todos os sistemas de aprendizado de máquina no primeiro degrau, o da associação. Eles são mestres na identificação de correlações, mas carecem dos modelos causais necessários para ascender aos níveis de intervenção e contrafactual, que são pré-requisitos para o raciocínio genuíno (Pearl, 2018). O progresso, portanto, tem sido na sofisticação da imitação, tornando a distinção filosófica de Searle (1980) entre a sintaxe e a semântica mais crítica à medida que a performance da simulação se torna mais convincente.

Se, como a análise indica, o paradigma computacional baseado em máquinas de Turing é inerentemente incapaz de gerar intencionalidade intrínseca, a sua emergência num substrato artificial exige uma ruptura radical com esse modelo. A investigação revelou duas propostas teóricas que articulam tal ruptura, partindo da premissa comum de que sistemas com "condições de contorno fixas" não podem resolver o Problema da Intrinsicalidade (Poznanski *et al.,* 2023). Contudo, essas propostas apontam para direções fundamentalmente opostas, definindo uma bifurcação no futuro da pesquisa.

O primeiro caminho, proposto por Penrose (1989), busca a solução nas leis mais fundamentais do cosmos. Argumenta que a consciência, evidenciada pela capacidade não algorítmica de compreensão matemática, depende de um processo físico não computável. Esse processo, a Redução Objetiva Orquestrada (Orch OR), seria um fenómeno quântico-gravitacional que ocorre nos microtúbulos neuronais, regido por uma Teoria da Gravitação

Quântica Correta (TGQC) ainda por descobrir. Nessa visão, a intencionalidade não é um privilégio da biologia, mas uma propriedade de uma lei física universal que a evolução biológica aprendeu a aproveitar.

O segundo caminho, articulado por Poznanski e colaboradores (2023), busca a solução na replicação dos princípios organizacionais da matéria viva. Propõe um hardware biomimético, um "*wetware* protônico", projetado para operar com "condições de contorno mutáveis" e auto-organização, características definidoras dos sistemas vivos. A intencionalidade emergiria de um processo de redução de incerteza baseado em dinâmica energética, e não em processamento simbólico. Nessa perspectiva, a intencionalidade está intrinsecamente ligada à arquitetura funcional da vida: a termodinâmica de não equilíbrio e a capacidade de um sistema definir as suas próprias restrições. Esse confronto não é apenas técnico, mas representa uma clivagem filosófica sobre a natureza da mente: é ela uma manifestação de uma lei física universal ou uma propriedade emergente da organização única que define a vida?

A jornada empreendida nesta dissertação – desde a materialidade da intencionalidade revelada pela arqueologia paleolítica até o impasse causal da IA algorítmica – obriga a uma reformulação da sua questão central. A investigação arqueológica ancorou a intencionalidade na sua história evolutiva, conferindo peso empírico à tese do naturalismo biológico sobre os "poderes causais do cérebro" (Searle, 1980). Este capítulo demonstrou que a IA atual está causalmente desligada dessa história. As propostas de Penrose e Poznanski representam, portanto, duas tentativas de criar uma *nova origem causal* para uma entidade artificial: uma fundada nas leis do cosmos, a outra nos princípios da vida.

Assim, a questão deixa de ser simplesmente *se* a intencionalidade poderá emergir em sistemas não biológicos, para se tornar *sob qual paradigma fundamental* a sua emergência poderia ser concebida. O problema da intencionalidade artificial é transformado de uma questão de engenharia incremental para uma de revolução científica fundamental. A resposta à pergunta "A intencionalidade permanecerá um atributo irredutível da consciência biológica ou poderá emergir em sistemas artificiais?" depende, em última análise, de qual desses dois monumentais e mutuamente exclusivos programas de pesquisa – o físico-universal ou o biomimético-funcional – se revelará, se algum deles o fizer, viável. Este capítulo conclui não com uma

resposta definitiva, mas com a delimitação clara dos horizontes para onde a investigação futura deve ser direcionada.

5 CONCLUSÃO

A trajetória percorrida nesta dissertação partiu de uma questão fundamental: a intencionalidade, a capacidade da mente de se direcionar ao mundo, permanecerá um atributo irredutível da consciência biológica ou poderá emergir em sistemas artificiais? Para construir uma resposta fundamentada, a pesquisa empreendeu uma jornada em três atos, partindo da definição filosófica da intencionalidade, atravessando sua gênese material na pré-história e, por fim, confrontando o desafio imposto pela Inteligência Artificial contemporânea.

O alicerce filosófico que sustentou toda a argumentação foi o naturalismo biológico de John Searle. Essa abordagem postula que a intencionalidade não é uma abstração ou uma metáfora, mas uma propriedade biológica real, *causada por* e *realizada na* estrutura neurofisiológica do cérebro. A distinção searleana entre a intencionalidade intrínseca da mente e a intencionalidade derivada de artefatos forneceu o critério ontológico e causal indispensável para a análise. Um estado intencional, para ser genuíno, depende não apenas de uma rede de outras crenças e desejos, mas, crucialmente, de um *background* de capacidades pré-intencionais e corporificadas, um saber-fazer que ancora o significado na interação com o mundo.

A investigação arqueológica, segundo ato desta jornada, ofereceu um contraponto empírico a essa base teórica, revelando a profunda historicidade da mente. Ao analisar o registro lítico através da metodologia da *chaîne opératoire*, foi possível rastrear a evolução da intencionalidade desde suas manifestações mais simples no *Olduvaiense*, que já evidenciavam a distinção entre intenção prévia e intenção-na-ação, até a complexa intencionalidade hierárquica e prospectiva exigida para a confecção de um biface Acheulense. A pesquisa demonstrou que a padronização e a transmissão dessas técnicas ao longo de vastos períodos de tempo seriam inexplicáveis sem a emergência de uma dimensão social da cognição, a intencionalidade conjunta de que fala Tomasello (2014), que funcionou como o mecanismo para a construção do *background* e da rede culturais. Esse processo foi impulsionado por um ciclo de retroalimentação positiva entre a cultura (ferramentas), a dieta (Hipótese do Tecido Caro) e a biologia (a evolução cerebral), conforme sustentado por autores como Aiello e

Wheeler (1995) e Holloway (1981), e cujos correlatos neurais são investigados pela neuroarqueologia de pesquisadores como Dietrich Stout. A intencionalidade humana, portanto, revelou-se não como uma propriedade abstrata, mas como um fenômeno profundamente corporificado, socialmente embebido e com uma história coevolutiva contingente.

O terceiro ato confrontou essa origem biológica com os desenvolvimentos de Sistemas Artificiais. A investigação de modelos e *frameworks* baseados em computação clássica, desde a IA simbólica até os Grandes Modelos de Linguagem, convergiu para uma conclusão unívoca: os sistemas atuais operam exclusivamente no domínio da intencionalidade derivada. Eles simulam com crescente fidelidade os efeitos do comportamento intencional, mas não replicam suas causas. Como argumentado por Kaufman (2022), a intencionalidade desses sistemas é um "eco estatístico" da intencionalidade coletiva humana presente nos dados de treinamento. O enquadramento da Escada da Causalidade de Judea Pearl (2018) formaliza esse limite, situando a IA atual no degrau da associação, incapaz de ascender aos níveis de intervenção e contrafactual que caracterizam o raciocínio genuíno. O argumento do Quarto Chinês de Searle (1980) permanece, assim, pertinente: a manipulação sintática, por mais sofisticada que seja, não gera compreensão semântica.

Com base na análise das barreiras fundamentais – a computacional e a biocausal – e no contraste radical entre a gênese encarnada da mente humana e a natureza desencarnada de Sistemas Artificias, a resposta à questão central desta dissertação é que, por enquanto, a intencionalidade intrínseca permanece como uma propriedade irredutível da consciência biológica e natural.

A expressão "por enquanto" é deliberada e crucial. A ciência é um processo exploratório, e as próprias pesquisas de fronteira indicam que a impossibilidade de uma intencionalidade artificial não é lógica, mas empírica e tecnológica no estágio atual. As investigações teóricas de Roger Penrose e R.R. Poznanski, exploradas nesta dissertação, abrem duas frentes de pesquisa que desafiam o paradigma computacional vigente. A proposta de Penrose (1989) de que a consciência depende de um processo físico não computável, a Redução Objetiva Orquestrada, aponta para uma revolução na física fundamental como pré-requisito para a mente artificial. Por outro lado, a via de Poznanski e colaboradores (2024) busca replicar os princípios funcionais da vida – como a auto-organização e as "condições de contorno mutáveis" – em um *hardware* biomimético, um "*wetware* protônico". Paradoxalmente, a própria IA, como

ferramenta de simulação e modelagem, pode se tornar um instrumento indispensável na investigação da física e da neurobiologia que poderiam, eventualmente, conduzir a uma intencionalidade não biológica.

Mesmo que a intencionalidade intrínseca não seja alcançada, a sociedade já lida com a onipresença de sistemas com agência real e intencionalidade simulada. A tendência humana ao antropomorfismo acarreta riscos de confiança mal colocada e manipulação, exigindo o desenvolvimento de um novo arcabouço sociocognitivo e ético para a interação com essas tecnologias.

Em última análise, a jornada da intencionalidade, desde a primeira pedra lascada até o algoritmo, revela o peso da história e a natureza do ser. A mente humana não é um programa desencarnado, mas o resultado de uma longa e complexa dança coevolutiva entre genes, matéria e cultura. A questão sobre se uma nova forma de intencionalidade poderá emergir em um substrato de silício permanece em aberto, mas a investigação aqui conduzida sugere que, se isso vier a ocorrer, não será por uma mera extrapolação da tecnologia atual, mas por uma ruptura fundamental que nos force a redefinir não apenas a máquina, mas a própria natureza da mente.

Deixo como uma provocação final a questão do fardo da intencionalidade. A mesma intencionalidade que um dia guiou a mão de um *hominínio* para lascar a primeira pedra deve, agora, com uma urgência e um discernimento infinitamente maiores, guiar a nossa mão para codificar o futuro.

A questão final que esta era nos coloca, não é se as máquinas se tornarão algum dia verdadeiramente intencionais. É se nós, como espécie, provaremos ser intencionais o suficiente para construir um futuro que valha a pena partilhar com elas.

REFERÊNCIAS BIBLIOGRÁFICAS

AIELLO, L. C.; WHEELER, P. The Expensive-Tissue Hypothesis: The Brain and the Digestive System in Human and Primate Evolution. **Current Anthropology**, Chicago, v. 36, n. 2, p. 199-221, abr. 1995.

BANDURA, A. **Social foundations of thought and action**: a social cognitive theory. Englewood Cliffs: Prentice-Hall, 1986.

BEAUNE, S. A. The invention of technology. **Current Anthropology**, Chicago, v. 45, n. 2, p. 139-162, abr. 2004.

BISSO-MACHADO, R. *et al*. Coevolução gene-cultura, teoria do nicho de construção e psicologia. **Psicologia-Reflexão e Crítica**, v. 35, n. 1, p. 1-10, 2022.

BOYD, R.; RICHERSON, P. J. **The origin and evolution of cultures**. New York: Oxford University Press, 2005.

BRATMAN, M. E. **Intention, plans, and practical reason**. Cambridge: Harvard University Press, 1987.

BRENTANO, Franz. **Psicologia do ponto de vista empírico**. 1. ed. São Paulo: Martins Fontes, 2006. (Tradução da obra original de 1874).

BYRNE, R. W. The social function of intellect. *In*: BYRNE, R. W.; WHITEN, A. (ed.). **Machiavellian intelligence:** social expertise and the evolution of intellect in monkeys, apes, and humans. Oxford: Clarendon Press, 1988.

CANAL, R. Resenha de "Mind: A Brief Introduction" de John R. Searle. **Princípios: Revista de Filosofia (UFRN)**, v. 13, n. 19-20, p. 283-290, 2006.

CORBEY, R. *et al.* The acheulean handaxe: more like a bird's song than a beatles' tune? **Evolutionary Anthropology: Issues, News, and Reviews**, v. 25, n. 1, p. 6-19, 2016.

CÓRDOBA, M. A. *et al.* Analyzing Intentional Behavior in Autonomous Agents under Uncertainty. *In*: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 32., 2023, Macau. **Anais** [...]. Macau: IJCAI, 2023. p. 6113-6121

COZMAN, F. G.; KAUFMAN, D. **Inteligência artificial**: uma abordagem filosófica. São Paulo: Livraria da Física, 2022.

CRAWFORD, K. **Atlas of AI**: Power, Politics, and the Planetary Costs of Artificial Intelligence. New Haven: Yale University Press, 2021.

DENNETT, D. C. **The intentional stance**. Cambridge: MIT Press, 1987.

DENNETT, D. C. A intencionalidade original. *In*: DENNETT, D. C. **A perigosa ideia de Darwin**: a evolução e os significados da vida. Rio de Janeiro: Rocco, 1994. p. 104.

DREYFUS, H. L. **What computers can't do:** a critique of artificial reason. New York: Harper & Row, 1972.

EGER, M.; MARTENS, C. Practical Specification of Belief Manipulation in Games. In: **AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE AND INTERACTIVE DIGITAL ENTERTAINMENT**, 13., 2017, Snowbird, EUA. Anais. Palo Alto: AAAI Press, 2017. p. 30-36.

FOXABBOTT, J. *et al*. Higher-Order Belief in Incomplete Information MAIDs. *In*: INTERNATIONAL CONFERENCE ON AUTONOMOUS AGENTS AND MULTIAGENT SYSTEMS, 23., 2024, Auckland. **Anais...** Auckland: IFAAMAS, 2024.

FLORIDI, L. AI as Agency without Intelligence. **Philosophy & Technology**, v. 37, n. 2, 2024.

FREITAS, M. C. S. A sociedade e a cultura da tecnologia. **Revista Tecnologia e Sociedade**, Curitiba, v. 10, n. 20, p. 1-17, 2014.

FRISTON, K. The free-energy principle: a unified brain theory? **Nature Reviews Neuroscience**, v. 11, n. 2, p. 127-138, 2010.

GALHARDO, D. Chaîne Opératoire: um método de análise para a compreensão da tecnologia lítica. **Revista do Museu de Arqueologia e Etnologia**, São Paulo, n. 25, p. 165-180, 2015.

GANASCIA, J. G. **L'intelligence artificielle**. Paris: Flammarion, 1996.

GOULD, S. J. *et al*. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. **Anais of the Royal Society of London.** Series B. Biological Sciences, v. 205, n. 1161, p. 581-598, 1982.

GREENFIELD, P. M. Language, tools and brain: the ontogeny and phylogeny of hierarchically organized sequential behavior. **Behavioral and brain sciences**, v. 14, n. 4, p. 531-551, 1991.

GRINDROD, J. Do LLMs (really) use language? **Minds and Machines**, v. 34, n. 1, p. 1-24, 2024.

HAMMOND, L.; EVERITT, T.; ABATE, A.; WITTE, E**.** On Imperfect Recall in Multi-Agent Influence Diagrams. In: CONFERENCE ON THEORETICAL ASPECTS OF RATIONALITY AND KNOWLEDGE – TARK, 19., 2023, Oxford, Reino Unido. **Anais**. Oxford: TARK, 2023. p. 123-134.

HAMEROFF, S.; PENROSE, R. Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness. **Mathematics and Computers in Simulation**, v. 40, n. 3-4, p. 453-480, 1996.

HARNAD, S. The symbol grounding problem. **Physica D: Nonlinear Phenomena**, v. 42, n. 1-3, p. 335-346, 1989.

HAUGELAND, J. **Artificial intelligence**: the very idea. Cambridge, MA: MIT Press, 1985.

HEYLGHIEN, F. The Global Superorganism: an evolutionary-cybernetic model of the emerging network society. **Social Evolution & History**, v. 6, n. 1, p. 58-119, 2007.

HOLLOWAY, R. L. Culture, symbols, and human brain evolution: A synthesis. **Dialectical anthropology**, v. 5, n. 4, p. 287-303, 1981.

HU, S.; CLUNE, J. Thought cloning: learning to think while acting by imitating human thinking. In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NeurIPS 2023), 37., 2023, New Orleans. **Anais**... [S.l.]: NeurIPS, 2023. Disponível em: https://github.com/ShengranHu/Thought-Cloning. Acesso em: 05 mar. 2025.

JHA, S.; RUSHBY, J. Inferring and conveying intentionality: beyond numerical rewards to logical intentions. In: AAAI Spring Symposium: Towards Conscious AI Systems, 2019, Stanford. **Anais** [...]. Palo Alto: AAAI Press, 2019. Disponível em: https://arxiv.org/abs/2207.05058. Acesso em: 14 fev. 2025.

KAUFMAN, D. Desmistificando a inteligência artificial: um guia para leigos e iniciados. São Paulo: Editora Blucher, 2022.

KAUFMAN, D.; SANTAELLA, L. A quarta ferida narcísica. **Estudos Avançados**, v. 38, n. 110, p. 7-22, 2024.

LALAND, K. N. Niche construction, evolution and culture. *In*: LALAND, K. N.; BROWN, G. R. (ed.). **Sense and nonsense:** evolutionary perspectives on human behaviour. Oxford: Oxford University Press, 2000.

LEMONNIER, P. **Elements for an anthropology of technology**. Ann Arbor: University of Michigan Press, 1992.

LEROI-GOURHAN, A. **Le geste et la parole:** technique et langage. Paris: Albin Michel, 1964.

LI, J. *et al.* **Exploring the Effects of Chatbot Anthropomorphism and Human Empathy on Human Prosocial Behavior Toward Chatbots**. arXiv preprint, arXiv:2506.20748, 2025. Disponível em: https://arxiv.org/abs/2506.20748. Acesso em: 09 mar. 2025.

LINDSAY, R. K. *et al*. **Applications of artificial intelligence for organic chemistry**: the DENDRAL project. New York: McGraw-Hill, 1980.

LYCETT, S. J. Heuristic approaches to the study of cultural transmission in the Lower Palaeolithic: The case of the Oldowan. **Quaternary International**, v. 355, p. 11-20, 2015.

LYRA, C. E. S.; MOGRABI, G. J. C.; EL-HANI, C. N. O naturalismo biológico e a relação entre consciência e cognição na filosofia de John Searle**. Trans/Form/Ação,** Marília, v. 39, n. 2, p. 143-162, 2016.

MAEDA, T. **Walkthrough of Anthropomorphic Features in AI Assistant Tools**. *arXiv preprint*, arXiv:2502.16345, 2025. Disponível em: https://arxiv.org/abs/2502.16345. Acesso em: 09 mar. 2025.

MORGAN, T. J. H. *et al*. Experimental evidence for the co-evolution of hominin tool-making teaching and language. **Nature Communications**, v. 6, n. 1, p. 6029, 2015.

MORENO DE SOUSA, J. C. **As pedras e os homens**: tecnologia lítica e cognição no povoamento do sul do Brasil. Curitiba: Appris, 2019.

NEGRANDO. A essência do aprendizado de máquina. **Nei Grando - Data & Analytics**, 4 maio 2022. Disponível em: https://neigrando.com/2022/05/04/a-essencia-do-aprendizado-de-maquina/. Acesso em: 05 mar. 2025.

NEVES, W. A. **O povo de Luzia**: em busca dos primeiros americanos. São Paulo: Globo Livros, 2020.

NUNES, Everardo P. **Tecnologia e educação**: o novo ritmo da informação. 2. ed. Campinas: Papirus, 2007.

PARK, P. S.; GOLDSTEIN, S.; O'GARA, A.; CHEN, M.; HENDRYCKS, D. AI deception: A survey of examples, risks, and potential solutions. **Patterns**, v. 5, art. 100988, may 2024.

PEARL, J.; MACKENZIE, D. **The Book of Why**: The New Science of Cause and Effect. New York: Basic Books, 2018.

PENROSE, R. **The emperor's new mind**: Concerning computers, minds, and the laws of physics. Oxford: Oxford University Press, 1989.

PENROSE, R. **Consciousness and the foundations of physics**. Cambridge: Cambridge University Press, 2023.

PENROSE, R.; HAMEROFF, S. Quantum computation in brain microtubules? The Penrose–Hameroff 'Orch OR' model of consciousness. **Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences**, v. 356, n. 1743, p. 1869-1896, 1998.

PLUMMER, T. W. *et al*. Expanded geographic distribution and dietary strategies of the earliest Oldowan hominins and Paranthropus. **Science**, v. 379, n. 6632, p. 561-566, 10 fev. 2023.

POZNANSKI, R. R. *et al.* The physics of life and the intrinsicality problem for AI. **Progress in Biophysics and Molecular Biology**, v. 183, p. 1-11, 2023.

POZNANSKI, R. R. *et al*. Intentionality for better communication in minimally conscious AI design. **Journal of Multiscale Neuroscience**, v. 3, n. 1, p. 1-12, 2024. Disponível em: https://doi.org/10.56280/1600750890. Acesso em: 12 jan. 2025.

PUTT, S. S. The functional brain networks that underlie Early Stone Age tool manufacture. **Nature Ecology & Evolution**, v. 1, n. 6, p. 1-8, 2017.

PYLYSHYN, Z. W. The role of connectionism in cognitive science: The story so far. **Ablex Series in Artificial Intelligence**, v. 2, p. 295-320, 1988.

RAO, A. S.; GEORGEFF, M. P. Modeling rational agents within a BDI-architecture. In: International Conference on Principles of Knowledge Representation and Reasoning – KR'91, 2., 1991, Cambridge. **Proceedings**… San Mateo: Morgan Kaufmann, 1991. p. 473-484.

RAO, A. S.; GEORGEFF, M. P. BDI agents: from theory to practice. In: International Conference on Multiagent System - ICMAS'95, 1., 1995, San Francisco. **Proceedings**... San Francisco: AAAI Press, 1995. p. 312-319.

REDAELLI, S. Preter-intentionality. The case of generative AI. **AI & SOCIETY**, p. 1-13, 2024.

RENFREW, C. The archaeology of religion. *In*: RENFREW, C.; ZUBROW, E. B. W. (ed.). **The ancient mind: elements of cognitive archaeology**. Cambridge: Cambridge University Press, 1994. p. 47-54.

RENFREW, C.; ZUBROW, E. B. W. (ed.). **The ancient mind: elements of cognitive archaeology**. Cambridge: Cambridge University Press, 1994.

RIBEIRO, A. Arqueologia da intenção. **Revista de Arqueologia**, v. 35, n. 1, p. 1-20, 2022.

RUSSELL, S.; NORVIG, P. **Artificial Intelligence**: A Modern Approach. 4. ed. Hoboken: Pearson, 2021.

SÁEZ, R. Una caja de herramientas de 2 millones de años. **Nutcrackerman**, 24 fev. 2015. Disponível em: https://nutcrackerman.com/2015/02/24/una-caja-de-herramientas-de-2-millones-de-anos/. Acesso em: 14 jun. 2025.

SÁNCHEZ OLSZEWSKI, S. A**. Designing Human-AI Systems:** Anthropomorphism and Framing Bias on Human-AI Collaboration. arXiv preprint, arXiv:2404.00634, 2024. Disponível em: https://arxiv.org/abs/2404.00634. Acesso em: 09 mar. 2025.

SANTONI DE SIO, F.; MECACCI, G. Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. **Philosophy & Technology**, v. 34, n. 4, p. 1057-1084, 2021.

SHORTLIFFE, E. H. **Computer-based medical consultations: MYCIN**. New York: Elsevier/North-Holland, 1976.

SEARLE, J. R. Minds, brains, and programs. **Behavioral and brain sciences**, v. 3, n. 3, p. 417-424, 1980.

SEARLE, J. R. **Intentionality**: an essay in the philosophy of mind. Cambridge: Cambridge University Press, 1983.

SEARLE, J. R. Consciousness, explanatory inversion, and cognitive science. **Behavioral and brain sciences**, v. 13, n. 4, p. 585-596, 1990.

SEARLE, J. R. **The rediscovery of the mind**. Cambridge: MIT press, 1992.

SEARLE, J. R. **The construction of social reality**. New York: Free Press, 1995.

SEARLE, J. R. **Consciousness and language**. Cambridge: Cambridge University Press, 2002.

SEARLE, J. R. **Mind**: A brief introduction. Oxford: Oxford University Press, 2006.

SEARLE, J. R. A mente. *In*: SEARLE, John R. **Mente, cérebro e ciência**. Lisboa: Edições 70, 2021. p. 1.

SHIPTON, C. *et al*. Acheulean biface variability in the eastern African Rift: a comparative study. **Journal of human evolution**, v. 125, p. 1-20, 2018.

SIMAS, C.; ULBRICHT, V. R**.** Human-AI Interaction: An Analysis of Anthropomorphization and User Engagement in Conversational Agents with a Focus on ChatGPT. In: INTERNATIONAL CONFERENCE ON APPLIED HUMAN FACTORS AND ERGONOMICS, 15., 2024, Nice, França. **Anais** [...]. New York: AHFE International, 2024.

SIMONDON, G. **On the mode of existence of technical objects**. Minneapolis: Univocal Publishing, 1980.

SIMONDON, G. **Sobre a técnica**. São Paulo: Ubu Editora, 2020.

SNYDER, J. M. *et al.* Oldowan technology was not transmitted by 'know-how'. **Scientific Reports**, v. 12, n. 1, p. 20739, 2022.

STOUT, D.; CHAMINADE, T. Stone tools, language and the brain in human evolution. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 367, n. 1585, p. 75-87, 2012.

STOUT, D. *et al*. Neural correlates of Early Stone Age toolmaking: technology, language and cognition in human evolution. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 363, n. 1499, p. 1939-1949, 2008.

TOMASELLO, M. **The cultural origins of human cognition**. Cambridge: Harvard university press, 1999.

TOMASELLO, M. **Origins of human communication**. Cambridge: MIT press, 2008.

TOMASELLO, M. **A natural history of human thinking**. Cambridge: Harvard University Press, 2014.

TOTH, N. The Oldowan reassessed: a close look at early stone artifacts. **Journal of Archaeological Science**, v. 12, n. 2, p. 101-120, 1985.

UOMINI, N. T.; MEYER, G. F. Shared brain lateralization patterns in language and tool use: a functional transcranial Doppler study. **PloS one**, v. 8, n. 8, p. e72693, 2013.

VAN ES, K.; NGUYEN, D. "Your friendly AI assistant": the anthropomorphic self-representations of ChatGPT and its implications for imagining AI. **AI and Society,** v. 40, n. 5, p. 3 591-3 603, 2025. DOI:10.1007/s00146-024-02108-6.

VASWANI, A. *et al.* Attention is all you need. *In*: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 30., 2017, Long Beach. **Anais** [...]. Long Beach: NIPS, 2017. p. 5998-6008.

WARD, F. R. *et al.* Honesty Is the Best Policy: Defining and Mitigating AI Deception. *In*: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 37, 2023, New Orleans. **Anais...** New Orleans: NeurIPS, 2023.

WARD, F. R. *et al.* The Reasons that Agents Act: Intention and Instrumental Goals. *In*: INTERNATIONAL CONFERENCE ON AUTONOMOUS AGENTS AND MULTIAGENT SYSTEMS, 23., 2024, Auckland. **Anais...** Auckland: IFAAMAS, 2024.

WEISS, G. (Ed.). **Multiagent Systems**: A Modern Approach to Distributed Artificial Intelligence. Cambridge, MA; London: The MIT Press, 1999. 619 p.

WORKDAY. **Learn about agentic AI**., 2024. Disponível em: https://www.workday.com/en-us/topics/ai/agentic-ai.html. Acesso em: 15 fev. 2024.

WYNN, T. Handaxe enigmas. **World Archaeology**, v. 27, n. 1, p. 10-24, 1995.

WYNN, T. Archaeology and cognitive evolution. **Behavioral and Brain Sciences**, v. 25, n. 3, p. 389-402, 2002.

WYNN, T. *et al*. Lomekwi 3: A new beginning or a false start? **Journal of Human Evolution**, v. 123, p. 108-118, 2018.