

**PONTIFÍCIA UNIVERSIDADE CATÓLICA DE SÃO PAULO  
PUC-SP**

**THIAGO GOMES DO NASCIMENTO**

**Confiança em agentes conversacionais:  
uma revisão narrativa**

Mestrado em Tecnologias da Inteligência e Design Digital

São Paulo  
2025

THIAGO GOMES DO NASCIMENTO

**Confiança em agentes conversacionais:  
uma revisão narrativa**

Mestrado em Tecnologias da Inteligência e Design Digital

Dissertação apresentada à Banca Examinadora da Pontifícia Universidade Católica de São Paulo, como exigência para obtenção do título de Mestre em Tecnologias da Inteligência e Design Digital, sob orientação do Prof. Dr. Diogo Cortiz da Silva.

São Paulo

2025

## BANCA EXAMINADORA

---

---

---

*O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Código de Financiamento 001.*

*This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Finance Code 001*

## **AGRADECIMENTOS**

Aos meus pais, exemplos de dedicação e comprometimento, e à minha família pelo apoio e incentivo que me fizeram acreditar que eu conseguiria superar todos os obstáculos ao longo da trajetória até aqui.

Ao meu orientador, Prof. Dr. Diogo Cortiz, pela confiança e generosidade partilhada, guiando com sua experiência e rigor acadêmico o aprimoramento deste trabalho.

À PUC-SP, em particular ao TIDD, todo seu corpo docente e discente pelas discussões enriquecedoras, que trouxeram profundidade e novas perspectivas à pesquisa.

À banca de qualificação, composta pela Prof.<sup>a</sup> Dra. Dora Kaufman e pelo Prof. Dr. Gabriel Rego pelas valiosas contribuições que ajudaram a tornar este trabalho mais maduro e robusto.

A todos os amigos, pelo apoio e pelas reflexões durante essa jornada, compartilhando os desafios e conquistas de cada etapa.

À CAPES, pelo financiamento e oportunidade.

NASCIMENTO, Thiago Gomes do. **Confiança em agentes conversacionais: uma revisão narrativa**. 2025. 81p. Dissertação (Mestrado em Tecnologias da Inteligência e Design Digital) – Pontifícia Universidade Católica de São Paulo, São Paulo.

## RESUMO

Com a evolução da Inteligência Artificial Generativa, surgem novas dinâmicas na interação com tecnologias, especialmente com agentes conversacionais inteligentes. Dada a importância da confiança na aceitação e uso dessas tecnologias, este trabalho tem como objetivo estudar os critérios que contribuem para a avaliação dos agentes. O desenvolvimento da fluência na linguagem trouxe as interações a um novo patamar, no qual as fronteiras entre o humano e o artificial são suavizadas. Assim, com a antropomorfização, os princípios da psicologia social passam a ser aplicados às máquinas, tendo como consequência o excesso de confiança na tecnologia. Para desenvolver uma avaliação adequada, são discutidas estratégias de calibração de confiança que envolvem o esforço cognitivo e o estilo de comunicação adotado.

**Palavras-chave:** inteligência artificial generativa; interação humano-computador; confiança; percepção social.

## ABSTRACT

Alongside to the widespread of Generative Artificial Intelligence, new dynamics of human technology interaction emerge, especially with intelligent conversational agents. Given the importance of trust in the acceptance and use of these technologies, this work aims to study the criteria that contribute to agents' assessment. The development of fluency in language has brought interactions to a new level, in which the boundaries between the human and the artificial are blurred. Thus, with anthropomorphism, the principles of social psychology begin to be applied to machines, , resulting in overconfidence in technology. In order to develop an appropriate reliance, trust calibration strategies involving cognitive forcing and communication style are discussed.

**Keywords:** generative artificial intelligence; human-computer interaction; trust; social perception.

## LISTA DE FIGURAS

Figura 1 - Modelo de confiança de Mayer <i>et al.</i> (1995). .....	11
Figura 2 - ELIZA .....	17
Figura 3 - Evolução dos agentes conversacionais .....	18
Figura 4 - Classificação dos agentes conversacionais <b>Erro! Indicador não definido.</b>	
Figura 5 - Taxonomia do comportamento comunicativo .....	20
Figura 6 - Cronologia dos LLM's.....	21
Figura 7 - Interação humano-computador .....	23
Figura 8 - Sugestões sociais .....	25
Figura 9 - Taxonomia de sugestões sociais.....	26
Figura 10 - Incorporação verbal.....	27
Figura 11 - Comportamento de máquina .....	30
Figura 12 - Tay (Microsoft).....	32
Figura 13 - Modelo teórico da confiança interpessoal nas Organizações .....	34
Figura 14 - Fatores com impacto na confiança.....	35
Figura 15 - Processo de formação da confiança .....	<b>Erro! Indicador não definido.</b>
Figura 16 - Modelo da confiança sustentada.....	45
Figura 17 - Antropomorfismo .....	47
Figura 18 - Interface de chat antropomórfico.....	48
Figura 19 - Vale misterioso .....	49
Figura 20 - Modelo de conteúdo de estereótipo .....	51
Figura 21 - Percepção social de agentes .....	52
Figura 22 - Bias Map .....	53
Figura 23 - Calibração da confiança.....	55
Figura 24 - Classificação das estratégias de debiasing.....	58
Figura 25 - Etapas para estratégia de esforço cognitivo .....	59
Figura 26 - Framework para metacognição.....	60
Figura 27 - Exemplo de "desantropomorfização" .....	61
Figura 28 - Influência dos modelos mentais na confiança na IA.....	62
Figura 29 - Método de calibração linguística .....	63
Figura 30 - Experimento com marcadores epistêmicos.....	64
Figura 31 - Proposta de calibração.....	65

<b>Figura 32</b> - Informações prévias e estilo de comunicação.....	66
<b>Figura 33</b> - Interface Copilot (Microsoft) .....	66
<b>Figura 34</b> - Interface Gemini (Google) .....	67
<b>Figura 35</b> - Interface ChatGPT (OpenAI) .....	68

## SUMÁRIO

<b>1. INTRODUÇÃO .....</b>	<b>10</b>
1.1. QUESTÃO E JUSTIFICATIVA DA PESQUISA.....	13
1.2. OBJETIVOS GERAIS E ESPECÍFICOS.....	14
1.3. HIPÓTESES .....	15
1.4. METODOLOGIA .....	15
<b>2. AGENTES CONVERSACIONAIS INTELIGENTES .....</b>	<b>16</b>
2.1. CONCEITO E EVOLUÇÃO .....	17
2.2. MODELOS DE LINGUAGEM .....	21
2.3. INTERAÇÃO HUMANO-COMPUTADOR.....	22
2.3.1. Computadores como atores sociais .....	24
2.3.2. Padrões enganosos .....	28
2.4. IA RESPONSÁVEL .....	29
<b>3. CONFIANÇA PERCEBIDA .....</b>	<b>33</b>
3.1. PROCESSO DE FORMAÇÃO DA CONFIANÇA.....	33
3.2. MÉTRICAS DE AVALIAÇÃO.....	38
3.3. ASPECTOS COGNITIVOS E SOCIAIS.....	39
3.4. JULGAMENTO E TOMADA DE DECISÃO .....	41
3.4.1. Viés de autoridade.....	41
3.4.2. Efeito Enquadramento .....	42
3.4.3. Atribuição causal .....	44
3.4.4. Antropomorfismo .....	46
3.4.5. Modelo de conteúdo de Estereótipos.....	49
<b>4. CONFIANÇA APROPRIADA.....</b>	<b>54</b>
4.1. CALIBRAÇÃO .....	54
4.2. ESTRATÉGIAS DE DEBIASING .....	57
4.3. ESTILO DE INTERAÇÃO .....	60
<b>5. CONSIDERAÇÕES FINAIS .....</b>	<b>69</b>
<b>REFERÊNCIAS .....</b>	<b>73</b>

## 1. INTRODUÇÃO

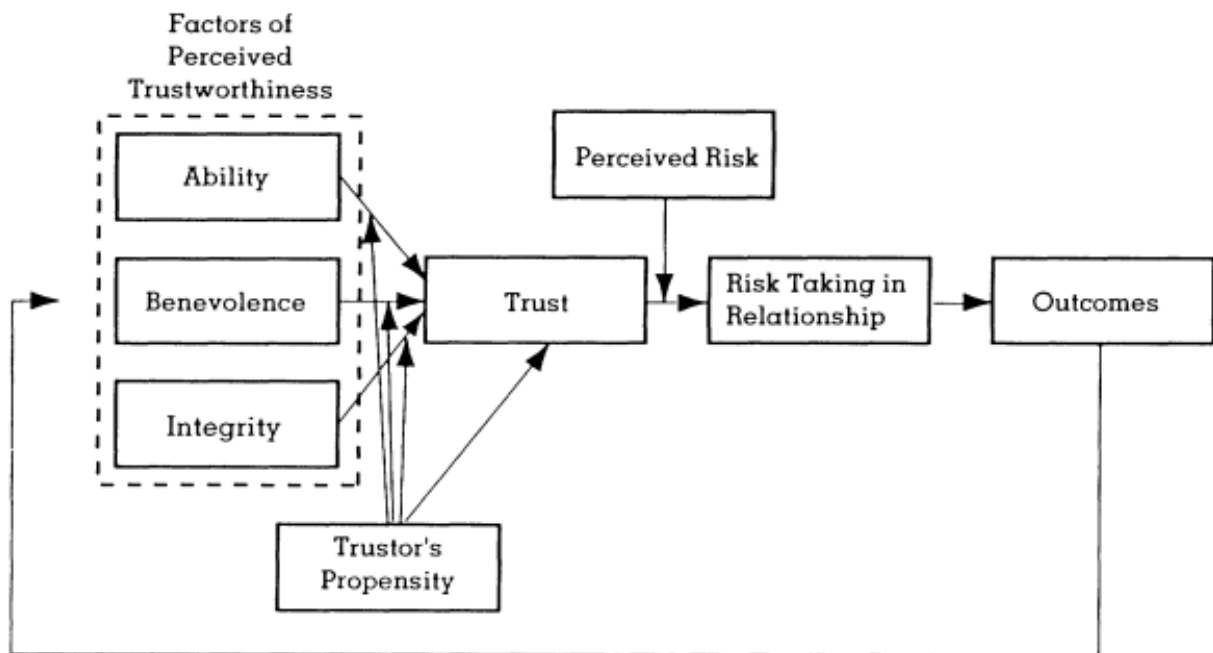
Em novembro de 2022, o lançamento do ChatGPT, modelo linguagem baseado em Inteligência Artificial (IA), projetado para interação em formato conversacional, alcançou grande repercussão midiática, deixando usuários impressionados com a sua capacidade de gerar parágrafos semelhantes ao conteúdo criado por humanos (Ji, 2024). A execução de tarefas de processamento de linguagem natural, como a geração de textos sintéticos aparentemente coerentes, foi impulsionada por inovações de arquitetura computacional aliadas ao crescimento do volume de dados de treinamento e da quantidade de parâmetros processados. Apesar dos interlocutores possuírem a tendência imputar significado e intenções às respostas recebidas, esses modelos operam a partir da combinação probabilística de padrões linguísticos, tendo sido denominados como meros papagaios estocásticos (Bender *et al.*, 2021).

As sugestões sociais incorporadas no estilo de comunicação utilizado pela tecnologia interativa associadas à antropomorfização, predisposição à atribuição de características humanas e estados mentais a agentes não humanos, podem interferir na confiança percebida pelos usuários, gerando expectativas inadequadas sobre a capacidade dos agentes (Feine *et al.*, 2019; Waytz; Cacioppo; Epley, 2010). Essa conexão remonta a um dos estudos clássicos da relação homem-máquina, no qual Muir (1987) define a confiança como uma função moderadora entre as características da máquina e a sua utilização pelo humano, pontuando a necessidade de se desenvolver a calibração dos julgamentos para o alcance de avaliações assertivas.

No contexto organizacional, Mayer et al. (1995) apresentam a confiabilidade percebida como uma composição dos fatores habilidade, benevolência e integridade, critérios que interagem entre si (Figura 1). Nessa dinâmica, ainda contribuem a propensão do avaliador a confiar em algo/alguém e o risco percebido. Amplamente replicado, o esquema foi retomado pelos autores em 2007 com a inclusão de aspectos referentes ao afeto, às emoções e aos padrões interculturais. Considerando perspectivas de diferentes áreas do conhecimento, a confiança está ligada a um relacionamento baseado na disposição em assumir riscos, em aceitar a vulnerabilidade (Schoorman; Mayer; Davis, 2007).

Ao explorar as relações sociais dos humanos à tecnologia, recorremos ao paradigma CASA (Computadores como atores sociais), que tem estimulado a realização de pesquisas experimentais acerca da interação humano-computador (IHC) com forma de evidenciar que os princípios da psicologia social, comunicação e sociologia, inicialmente propostos para as interações entre humanos, também são aplicáveis para explicar como as pessoas se relacionam com a tecnologia.

**Figura 1** - Modelo de confiança de Mayer *et al.*(1995).



Fonte: Mayer *et al.* (1995).

Estudos empíricos constatam que os vieses cognitivos, padrões de julgamento que levam a decisões irracionais, exercem grande influência na tomada de decisão dos usuários (Huang; Twidale; Bashir, 2019; Santhanam; Karduni; Shaikh, 2020). Portanto, a atribuição de confiança sofre influência de normas sociais que atuam no sentido de promover a colaboração a partir da percepção do comportamento do outro. Isso ocorre porque temos uma atitude afetiva otimista acerca da competência e das boas intenções do outro, além da perspectiva de que ele será impactado positivamente pelas crenças formadas. Em suma, há a previsão de uma resposta positiva resultante da interação, com base nas intenções percebidas, ainda que não haja evidência concreta para tal (Jones, 1996).

As heurísticas de confiança, por exemplo, são atalhos mentais acionados na avaliação de agentes inteligentes que simplificam julgamento e a realização de escolhas. Usuários tendem a considerar sinais de autoridade nas interações realizadas em ambientes digitais ou ainda desenvolver uma posição mais favorável à tecnologia tendo como referência experiências anteriores. Mesmo sendo um processo natural, as heurísticas podem ser utilizadas intencionalmente no desenvolvimento dos sistemas como estratégia para estimular sua adoção, promovendo a aceitabilidade dos resultados gerados (Liao; Sundar, 2022).

Na avaliação de agentes conversacionais, especificamente, Rheu *et al.* (2021) apresentam um conjunto fatores capazes de aumentar o nível de confiança atribuído pelos usuários. O primeiro deles diz respeito à inteligência socioemocional do agente, identificada através do diálogo e de expressões, inclusive sobre suas deficiências de desempenho. Por meio do estilo de comunicação e características da voz utilizada (neste caso, com uma ferramenta que exibe características vocais), o usuário pode encontrar similaridades com os estereótipos com os quais está habituado. Outro fator presente em agentes incorporados, aqueles que possuem uma representação física, é a aparência antropomórfica do agente, que quando associada à comunicação não verbal, traz familiaridade para a experiência. O último tema está ligado à competência, seja quanto à qualidade dos resultados ou quanto à reputação do agente.

Em um estudo qualitativo quanto ao uso de agentes conversacionais em ambientes digitais de trabalho, Gkinko e Elbanna (2023) destacaram o componente emocional da confiança, associado a características similares a atributos humanos. Adicionalmente, Kim *et al.*(2023) evidenciaram a influência de fatores contextuais na avaliação de ferramentas de IA pelos usuários e a importância de estabelecer uma confiança calibrada, o que requer um entendimento mais amplo dos fatores que interferem na confiança e como eles interagem entre si.

Diante do exposto, fica evidente a necessidade de investigar com maior profundidade como se dá a avaliação dos agentes inteligentes. Esse entendimento poderá ser utilizado como embasamento para ações focadas na redução das distorções, já apontadas em estudos anteriores, entre a confiança percebida e desempenho real dos agentes, que levam tanto ao excesso de confiança (Passi; Vorvoreanu, 2022) quanto à aversão algorítmica (Dietvorst; Simmons; Massey, 2015).

Nas próximas seções, esses itens serão abordados com maior detalhe, a fim de trazer os conceitos e processos presentes no contexto do problema de pesquisa. Tendo delimitado o escopo deste trabalho, resta explicitar a pergunta norteadora que guiará as discussões levantadas.

## 1.1. QUESTÃO E JUSTIFICATIVA DA PESQUISA

Tendo em vista o contexto apresentando, este trabalho pretende responder a seguinte questão de pesquisa: “Como os usuários avaliam a confiabilidade dos agentes conversacionais inteligentes?”.

Em virtude das fragilidades inerentes aos agentes conversacionais inteligentes, torna-se imperativo o estudo de abordagens que promovam seu uso de forma responsável (Petrov *et al.*, 2023). Em linha com a posição defendida por Goddard *et al.* (2012), este trabalho visa promover a comunicação das limitações dos modelos. Para isso, resgatamos o convite realizado por Nass e Moon (2000) em direção ao desenvolvimento de estudos capazes de compreender de maneira granular como as características dos sistemas são recebidas por diferentes perfis de usuários, sobretudo em um momento de profusão de novas ferramentas e modos de interação.

Como resposta à recorrente falha de utilização de medidas de confiança de forma monolítica e personalizada, sem uma profunda análise dos instrumentos já desenvolvidos, este trabalho contempla a complexidade e multiplicidade de dimensões da confiança, compilando as contribuições disponíveis sobre o tema, como meio de identificação de aspectos específicos para futuras investigações (Kohn, 2023).

A visão interdisciplinar promovida permite captar os padrões de comportamento integrando psicologia social e inteligência artificial. Mesmo que o contexto dos agentes conversacionais baseados em modelos generativos de linguagens possua suas particularidades, o conhecimento construído em áreas de conhecimento correlatas, seja na interação social ou na interface com outros tipos de tecnologias podem ser úteis para a compreensão da dinâmica da confiança e suas consequências.

Além disso, conforme sinalizado por Chiou e See (2023), há uma predominância de estudos voltados ao aumento do grau de confiança atribuído à tecnologia. No campo da interação humano-computador, a aplicação de sugestões sociais responsáveis por evocar um maior engajamento com os agentes pode ser controversa nos cenários onde é estimulada a adoção de uma postura de risco mesmo diante de uma baixa competência e prejudicando a experiência do usuário (Weidinger *et al.*, 2021; Alberts *et al.*, 2024).

Portanto, como forma de preencher as lacunas identificadas e fornecer as contribuições teóricas e práticas das quais a área de conhecimento carece, este trabalho também vem ao encontro da necessidade latente de fornecer subsídios para o debate social da relação com a tecnologia, incluindo a exposição das estratégias cognitivas a serem aplicadas para uma avaliação assertiva dos agentes inteligentes, tendo como consequência uma menor suscetibilidade a erros e maior aderência dos sistemas às expectativas dos usuários.

## 1.2. OBJETIVOS GERAIS E ESPECÍFICOS

O objetivo deste trabalho é caracterizar os critérios associados à atribuição de confiança em agentes conversacionais inteligentes sob a perspectiva dos usuários, com foco nos vieses cognitivos evocados na percepção, julgamento e tomada de decisão, além de abordar os possíveis impactos de avaliações descalibradas e as mitigações aplicáveis. Como objetivos específicos, foram determinados os seguintes:

- a) Mapear as características dos agentes conversacionais inteligentes capazes de gerar confiança no âmbito da interação com os usuários.
- b) Identificar os mecanismos da cognição humana que influenciam a tomada de decisão no contexto da interação social e com demais agentes.
- c) Analisar dinâmica da confiança percebida a partir da combinação entre características dos usuários, dos agentes e fatores contextuais.
- d) Levantar recomendações de intervenções com potencial de gerar avaliações mais assertivas.

### 1.3. HIPÓTESES

A partir do levantamento prévio da literatura, foram traçadas as seguintes hipóteses:

H1: As sugestões sociais incorporadas as agentes conversacionais são capazes de evocar reações sociais nos usuários, como empatia e confiança.

H2: Processos heurísticos impactam as escolhas e julgamentos referentes tanto às interações interpessoais quanto às relações com agentes não humanos.

H3: As avaliações de confiança podem ser excessivamente elevadas, mesmo quando os agentes não possuem um nível de competência correspondente.

H4: Intervenções com apelo cognitivo são capazes de calibrar a confiança percebida pelos usuários.

### 1.4. METODOLOGIA

Este trabalho trata-se de uma revisão narrativa, que busca compreender a dinâmica da confiança em agentes conversacionais inteligentes, a fim de identificar os fatores que exercem influência nas avaliações realizadas pelos usuários, contribuindo com o levantamento de estudos de diferentes áreas do conhecimento para a sugestão de recomendações que direcionem para uma maior alinhamento entre a percepção e a capacidade real dos agentes.

As revisões narrativas são trabalhos qualitativos, que têm como objetivo descrever e discutir o estado da arte de um tema a partir de um perspectiva teórica ou contextual, consistindo em uma análise crítica da literatura disponível em livros e artigos de periódicos (Rother, 2007). Ao conectar diferentes estudos, a revisão narrativa consiste em um meio de identificar problemas, desenvolver teorias ou gerar novas hipóteses. Neste sentido, é possível abordar questões mais amplas, reunindo evidências advindas de estruturas metodológicas diversas (Baumeister; Leary, 1997).

Considerando a triangulação entre agente, humano e contexto, a primeira etapa da pesquisa se debruçará sobre a evolução dos agentes conversacionais até o estado atual da aplicação de modelos de linguagem generativos, destacando contribuições e riscos identificados no âmbito do framework CASA (Computadores como atores sociais), referência na área da interação humano-computador, segundo o qual os indivíduos aplicam regras sociais inconscientes na interação com as máquinas (Nass; Moon, 2000; Nass; Steuer; Tauber, 1994).

Em seguida, será aprofundada a discussão voltada aos usuários, explorando o quadro teórico da cognição social no tocante a percepção, julgamento e tomada de decisão e comportamento de escolha. Desta forma, serão expostas as interseções entre os conceitos de heurísticas e vieses (Tversky; Kahneman, 1974), o modelo de conteúdo de estereótipos (Fiske *et al.*, 2002) e estudos empíricos sobre aceitação e confiança na tecnologia.

Por fim, são sugeridas recomendações em direção ao projeto de agentes com um nível apropriado de confiança (Lee; See, 2004), com a utilização de estratégias de desvios e intervenções no estilo de comunicação adotado pelos agentes.

## **2. AGENTES CONVERSACIONAIS INTELIGENTES**

A Inteligência Artificial (IA) Generativa tem levantado debates quanto a sua capacidade de transformação da sociedade. Aplicações que permitem criar conteúdo personalizado já contribuem com a realização de tarefas em diferentes áreas.

Sistemas de IA Generativa criam conteúdo com base em dados de treinamento e em resposta a *prompts* de usuários. Seu crescimento recente e a cobertura da mídia destacaram as capacidades da IA, levando a discussões significativas no âmbito público, acadêmico e político. Além de criar conteúdo sintético, os sistemas de IA generativa estão sendo cada vez mais usados como agentes autônomos, permitindo que os modelos ultrapassem as datas-limite de seus dados de treinamento e adquiram novas possibilidades. A IA generativa tem o potencial de revolucionar indústrias e a sociedade, sendo já utilizada para criar conteúdo individualizado e escalável, automatizar tarefas e melhorar a produtividade (OECD, 2023, p. 8, tradução nossa).

Ao longo deste capítulo, serão identificadas as características dos agentes conversacionais inteligentes que os tornam objetos relevantes para o estudo da confiança, abordando não só a forma como eles operam como também as consequências advindas do seu modo de construção e utilização.

## 2.1. CONCEITO E EVOLUÇÃO

Agentes conversacionais são sistemas com os quais é possível interagir utilizando linguagem natural. Apesar de apresentados em nomenclaturas distintas como agentes virtuais, *chatbots* ou sistemas dialógicos, compartilham de um mesmo objetivo: simular a conversa humana, utilizando modelos de linguagem e algoritmos para emular a comunicação entre o usuário e o computador (Shawar & Atwell, 2007).

Ao resgatar a origem das interfaces conversacionais, o programa ELIZA (Weizenbaum, 1966) é considerado seminal na interação homem-máquina. Seu método de funcionamento consistia na identificação de palavras-chave nos dados de entrada, que atuavam como gatilhos para a transformação em uma resposta associada a regras pré-estabelecidas. Apesar do entusiasmo gerado na época, limitações técnicas restringiam seu desempenho.

**Figura 2 - ELIZA**

```

Welcome to
          EEEEE LL    IIII  ZZZZZ  AAAAA
          EE    LL    II    ZZ    AA  AA
          EEEEE LL    II    ZZ    AAAAAA
          EE    LL    II    ZZ    AA  AA
          EEEEE LLLLLL IIII ZZZZZ  AA  AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.

```

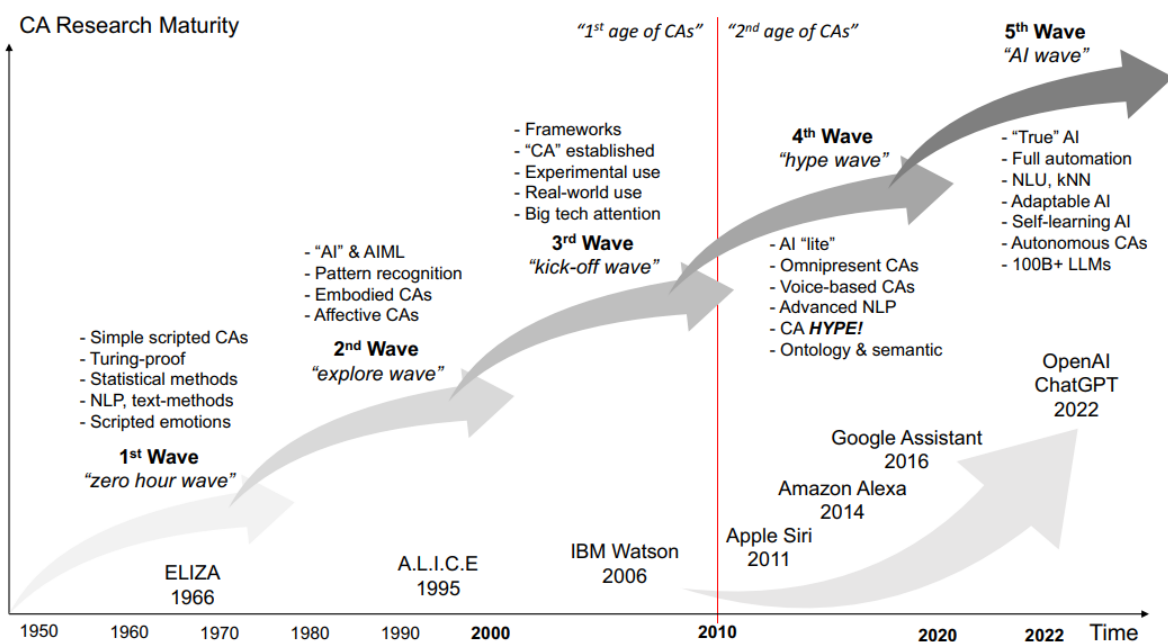
Fonte: Wikimedia Commons (2023).

Schöbel *et al.* (2024) apresentam cinco ondas de evolução, traçando o progresso dos agentes conversacionais desde o ELIZA até os dias atuais (Figura 3).

A segunda onda possuía um carácter exploratório, com o uso do reconhecimento de padrões e das primeiras soluções de inteligência artificial, tendo como principal representante o A.L.I.C.E., lançado em 1995 (WALLACE, 2009). Também foi nesse período que surgiram os primeiros agentes incorporados com características humanas, ainda rudimentares, marcando um distanciamento entre as pesquisas teóricas e os artefatos práticos.

Os desenvolvimentos ocorridos nos anos 2.000 deram início aos agentes conversacionais que se assemelham aos que estão disponíveis até agora. Termos como *chatbot* passaram a ser utilizados para designar agentes como o Watson (IBM), em 2006, que atraiu a atenção de outras grandes empresas do mercado de tecnologia. A partir de 2010, impulsionada pelo processamento de linguagem natural e pela IA, a quarta onda dá início a uma nova era dos agentes conversacionais. Esse marco representou a popularização das aplicações, inclusive com interações por voz. Nesse contexto, surgem assistentes virtuais como Siri (Apple), e Alexa (Amazon). Por fim, a quinta onda, destaca os agentes autônomos e adaptáveis viabilizados pelos grandes modelos de linguagem, que em 2022 ganharam as manchetes após o lançamento do ChatGPT(OpenAI) (SCHÖBEL *et al.*, 2024).

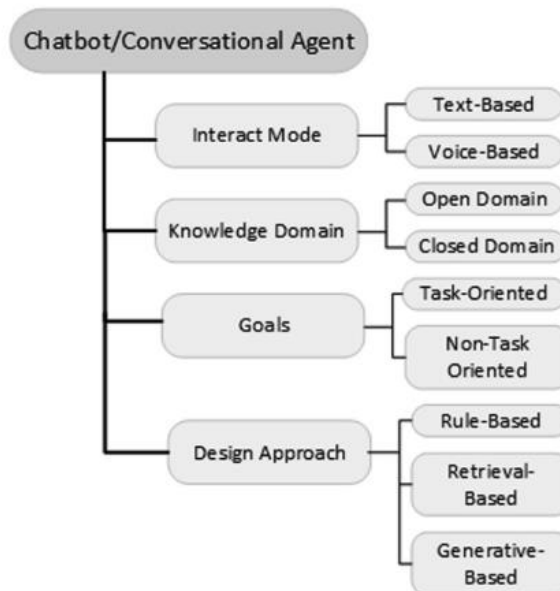
**Figura 3 - Evolução dos agentes conversacionais**



Fonte: Schobel et. al (2024)

Ao longo do tempo, novas formas de construção e interação foram desenvolvidas. Na classificação proposta por Hussain *et al.* (2019) a comunicação com os agentes pode ser baseada em texto ou voz. A depender do treinamento recebido, o conhecimento do *chatbot* pode ser restrito a determinada área, geralmente orientado a tarefas específicas, ou ter um caráter mais versátil, incluindo conversas com objetivo de gerar entretenimento. As técnicas utilizadas no desenvolvimento também são agrupadas categorias. Partindo das mais simples para as mais avançadas, temos as baseadas em regras, na recuperação de respostas em uma base pré-existente ou em modelos generativos, que detalharemos na seção seguinte. Outras classificações mencionam os agentes conversacionais incorporados, que por sua vez são aqueles que possuem uma representação gráfica, seja em 2 ou 3 dimensões. As aplicações destes agentes são diversas, passando pelas áreas de atendimento ao consumidor, jogos digitais, acessibilidade, assistentes pessoais, tutores para educação à distância e até terapeutas virtuais (Barros e Tedesco, 2016).

**Figura 4** - Classificação dos agentes conversacionais



Fonte: Hussain *et al.* (2019)

A naturalidade percebida nas interações não ocorre por acaso. Diversas características da comunicação humana são inseridas na concepção dos agentes. Após realizar uma extensa revisão da literatura, Van Pinxteren *et al.* (2020)

identificaram alguns padrões, criando uma taxonomia de comportamentos de comunicação. Os autores criaram uma matriz contendo a modalidade da interação no eixo vertical e os modelos mentais evocados no eixo horizontal. Dentre os quadrantes gerados, os componentes verbais baseados na responsividade são apresentados como os mais relevantes para os *chatbots*. Além do próprio estilo de comunicação, a demonstração de educação, humor e empatia aumentam o nível de utilização dos agentes.

**Figura 5 - Taxonomia do comportamento comunicativo**

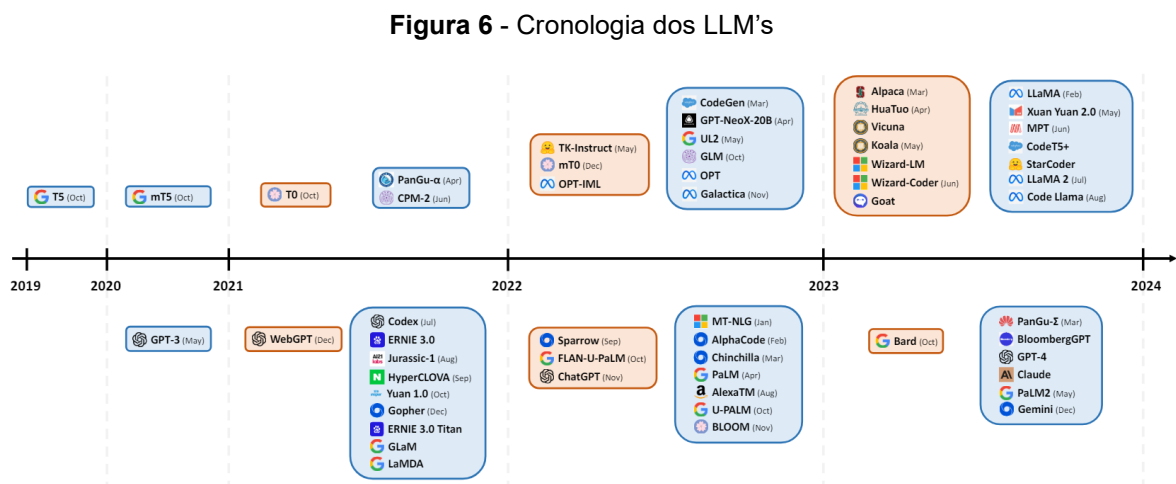
		Footing			
		Human similarity (26)	Individual similarity (15)	Responsiveness (26)	Other (3)
Modality	Nonverbal behaviors (26)	<b>Nonverbal behaviors grounded in human similarity</b> voice ● gaze ●● body movements ●● facial expressions and body movements ●●● gaze and body movements ● behavioral realism ● gaze and breathing ● gaze and head movements ● congruency of gestures ●●	<b>Nonverbal behaviors grounded in individual similarity</b> response time ● head movements ●● voice ● personality ●	<b>Nonverbal behaviors grounded in responsiveness</b> etiquette ● response time ● listening behaviors (e.g. nodding) ●● empathic facial expressions ● cooperative gestures ● laughter ●	<b>Other nonverbal behaviors</b>
	Verbal behaviors (25)	<b>Verbal behaviors grounded in human similarity</b> cognitive recall ●	<b>Verbal behaviors grounded in individual similarity</b> communication style ● personality ●	<b>Verbal behaviors grounded in responsiveness</b> affect support ● social praise ●● politeness ●●● humor ●●●● communication style ● small talk ●● empathy ● expectancy setting and service recovery ● cooperative utterances ● responsive utterances ● social role ●	<b>Other verbal behaviors</b> transparency ● argument type ● confidence cues ●
	Appearance Characteristics (19)	<b>Appearance characteristics grounded in human similarity</b> appearance ●●●●● facial appearance ●●● physical embodiment ●	<b>Appearance characteristics grounded in individual similarity</b> ethnicity ● gender and ethnicity ● physical appearance ● facial appearance ●● physical appearance and environment ● physical presence ●●	<b>Appearance characteristics grounded in responsiveness</b>	<b>Other appearance characteristics</b> attractiveness ●

Fonte: Van Pinxteren et al.(2020)

Dentre os fatores que estimularam o desenvolvimento dos agentes conversacionais, estão os avanços na IA, especialmente o aprendizado profundo, viabilizado pela crescente capacidade de processamento e alto volume de dados obtidos através da para treinamento dos modelos, obtidos como resultado da maior conectividade dos usuários, cadastrados em serviços de nuvem. Além disso, a evolução das tecnologias de do reconhecimento de discurso e processamento de linguagem natural tem despertado o interesse por aplicações com interfaces conversacionais como ferramenta de vantagem competitiva (McRae, 2016).

## 2.2. MODELOS DE LINGUAGEM

A linguagem está no cerne da comunicação e da racionalidade humana. A habilidade de transferir conhecimento por meio da linguagem é considerada um dos motivos de sucesso e evolução da espécie humana na natureza (LUPYAN, 2016). Conforme sinalizado na cronologia elaborada por Naveed *et al.* (2024), exibida na Figura 6, o desempenho dos modelos de linguagem foi beneficiado pelo desenvolvimento de código aberto aliado à aplicação de técnicas de refinamento.



Fonte: Naveed *et al.* (2024).

Fomentados pelo acesso a amplas bases de treinamento, os algoritmos de processamento de linguagem natural tiveram um salto de capacidade. Assim, foi possível agrupar atividades como respostas a perguntas, tradução, compreensão ou sumarização de textos em grandes modelos de linguagem. Com isso, os agentes virtuais puderam desenvolver a habilidade de conduzir diálogos de forma convincente (Radford *et al.*, 2019).

Os seres humanos são seres sociais acostumados à interação social na forma de diálogos baseados em perguntas e respostas. Essa interação com o robô torna-se mais intuitiva e natural do que uma operação fria e unilateral, característica de softwares. Os humanos estão habituados a se comunicar por meio de uma linguagem mútua. Essas características de engajamento em discussões com IA generativa, conectando ideias por meio de um exame mútuo de informações internas, conversas hipotéticas, diálogos envolventes e a ligação de princípios existentes a contextos mais atuais são mais intuitivamente corretas porque os humanos tendem a examinar logicamente as informações quando apresentadas de forma que correspondam à sua percepção de fenômenos ou a exemplos cotidianos (Obrenovic *et al.*, 2024, tradução nossa).

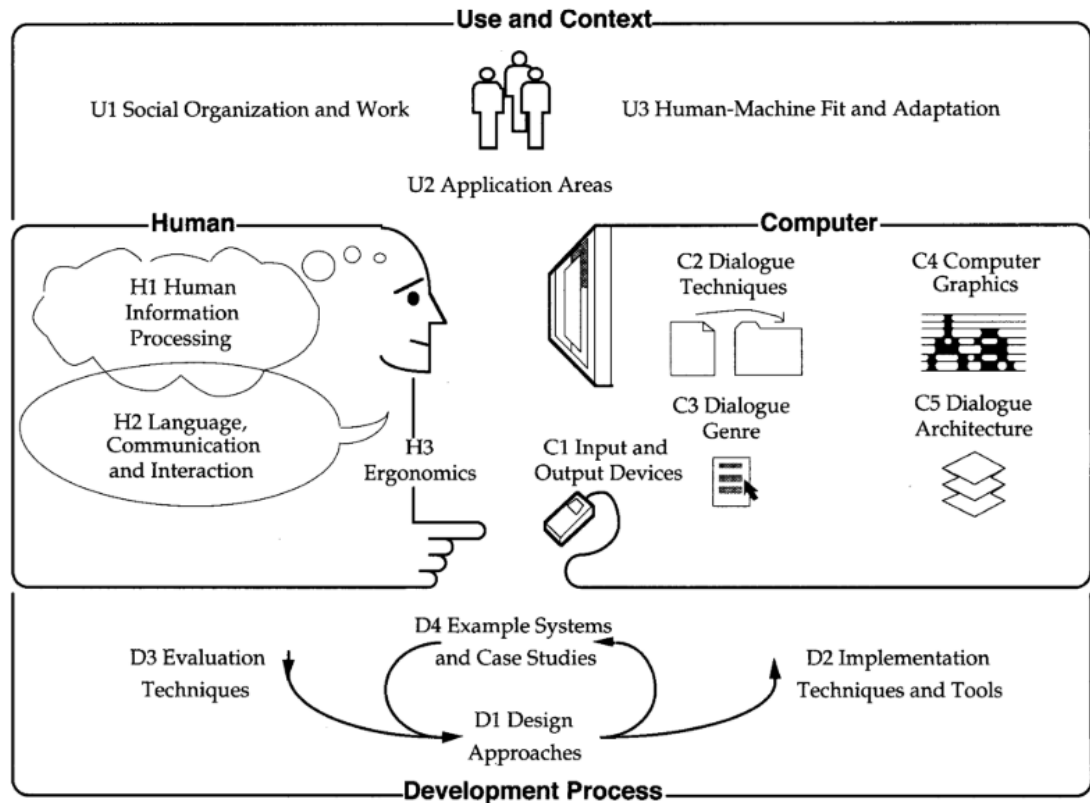
Em seu processamento, os modelos utilizam a técnica de tokenização, processo de divisão dos textos em palavras, subpalavras ou até caracteres, posteriormente codificados e vetorizados para que possam ser tratados estatisticamente. Com essa representação, é possível aplicar os algoritmos de *word embedding*, responsáveis por identificar as sequências de tokens que se combinam com maior frequência. Porém, é necessário observar que os resultados desse processamento carregam os vieses presentes nos dados e nas técnicas de refinamento aplicadas, gerando disparidades para línguas e comunidades subrepresentadas (Petrov *et al.*, 2023).

### 2.3. INTERAÇÃO HUMANO-COMPUTADOR

A interação humano-computador (IHC), como campo de estudo, envolve o projeto, avaliação e implementação de sistemas computacionais interativos para uso humano, além dos fenômenos significativos relacionados a eles. Dado seu caráter interdisciplinar, faz uso de conhecimentos advindos da ciência da computação, psicologia, sociologia, antropologia e *design*. Com esse ferramental, dedica-se a analisar o desempenho de atividades e a estrutura da comunicação entre humanos e máquinas, abarcando interfaces, algoritmos e especificações. A figura 7 ilustra como a IHC acontece em um contexto social, que de acordo com as áreas de aplicação passa por adaptações, tendo em vista as características dos usuários e as técnicas utilizadas (Hewett *et al.*, 1992).

A percepção do usuário é um fator crucial em todo processo de interação, inclusive quando se trata do diálogo com agentes conversacionais inteligentes. Ao abordar o conceito de *affordances*, cunhado por Gibson (1979) para denominar as possibilidades de ação que identificamos em relação ao ambiente, Norman (1988) estabeleceu uma nova interpretação focada no *design*. Por essa perspectiva, o usuário identifica a utilidade de um artefato analisando de forma conjunta as sugestões presentes em sua estrutura, o repertório pessoal e contexto ambiental. Portanto, as características consideradas na construção dos agentes são capazes de promover comportamentos específicos, podendo levar a uma maior propensão ao uso da tecnologia.

**Figura 7 - Interação humano-computador**



Fonte: Hewett *et al.* (1992)

As *affordances* sociais, em particular, são “possibilidades para interação ou possibilidades de ação que são moldadas por práticas e normas sociais” (Carvalho, 2020, p. 1) Relacionando a complexidade da interação com as *affordances*, podemos ver os simples movimentos corporais de nossos interlocutores como oportunidades para diferentes tipos de ações, assim como no ambiente físico. Por exemplo, um gesto específico acompanhado de um olhar pode ser interpretado como uma chance de se aproximar, se afastar, abraçar, beijar, empurrar ou sorrir para alguém. Da mesma forma, as vocalizações são percebidas como indicadores de possibilidades para novas ações verbais, como expressar algo de forma afirmativa, avaliativa, solidária, alegre, provocadora, flertiva ou antagônica. Gestos, expressões faciais, vocalizações, e outros comportamentos criam oportunidades para futuras interações no diálogo humano; eles convidam a novas ações (Jensen; Pedersen, 2016).

### 2.3.1. Computadores como atores sociais

Reeves & Nass (1996) evidenciaram que as pessoas tendem a responder socialmente à tecnologia, ou seja, aplicam o padrão de etiqueta nas interações, além de considerar traços de personalidade e desenvolver uma relação emocional com os dispositivos. Com base nisso, propuseram o paradigma do “Computador como ator social”, um *framework* para teste de hipóteses na área da interação humano-computador, composto pelas seguintes etapas:

**Quadro 1-** Framework CASA (Computador como ator social).

Item	Etapa
1	Escolha uma descoberta das ciências sociais (teoria e método) que trate de comportamentos ou atitudes em relação a humanos. Os estudos apresentados aqui baseiam-se na psicologia social e na sociologia.
2	Substitua "humano" por "computador" na formulação da teoria.
3	Troque um ou mais humanos por computadores no método do estudo.
4	Proporcione ao computador características associadas aos humanos: (a) resposta de linguagem; (b) respostas baseadas em múltiplas entradas anteriores; (c) desempenho de papéis tradicionalmente associados a humanos; e (d) produção de vozes com características humanas.
5	Determine se a regra social ainda se aplica.

Fonte: Reeves & Nass (1996)

Após reunir evidências de uma série de experimentos, Nass e Moon (2000) confirmaram que a atribuição de características sociais como gênero, etnia e personalidade aos computadores provocava uma série de expectativas, suposições e atribuições nos indivíduos, tendo como consequência percepções de similaridade, reciprocidade e até de confiança nas interações. Os participantes dos estudos apresentaram maior receptividade aos computadores mediante a inserção de sugestões suficientes para classificá-los como integrantes de um mesmo grupo ou dotados de um padrão de comportamento esperado.

Podemos concluir que os indivíduos estão reagindo de forma inconsciente aos computadores na medida em que aplicam normas sociais — normas para interações humano-humano — que são inadequadas para interações humano-computador, basicamente ignorando os sinais que revelam a natureza essencialmente não social de um computador. Na prática, descobrimos que o melhor meio de identificar essas regras é recorrer à literatura da psicologia social experimental (Nass & Moon, 2000, p.83, tradução nossa).

Ao declarar a necessidade de evolução com uma segunda geração do paradigma, Nass e Moon (2000) indicam que a questão estará plenamente respondida quando houver um entendimento mais granular de quais características dos computadores têm impacto sobre quais indivíduos, além do detalhamento dos motivos que levam à aplicação das regras sociais em uma maneira similar à interação entre humanos. Diversos autores se propuseram a contribuir com a continuidade proposta, com a replicação de estudos da psicologia social para a interação humano-computador. Ao longo deste trabalho essas contribuições serão analisadas.

Fogg (2003) acrescenta o atributo da persuasão à definição dos computadores como atores sociais, visto que eles capazes de influenciar decisões através da criação de relacionamento, fornecendo *feedback* positivo aos usuários, modelando o comportamento em direção a um padrão ou fornecendo suporte social. A interação social pode acontecer não só através da linguagem, mas também de forma psicológica, com a demonstração de empatia e preferências, de maneira dinâmica, como na adoção de uma postura cooperativa, através do papel social exercido pela aplicação (professor, médico, colega de trabalho) ou ainda de forma física, quando o agente apresenta características como face, olhos e movimentos corporais. Todos esses fatores podem interferir na avaliação de quão confiável determinado agente é.

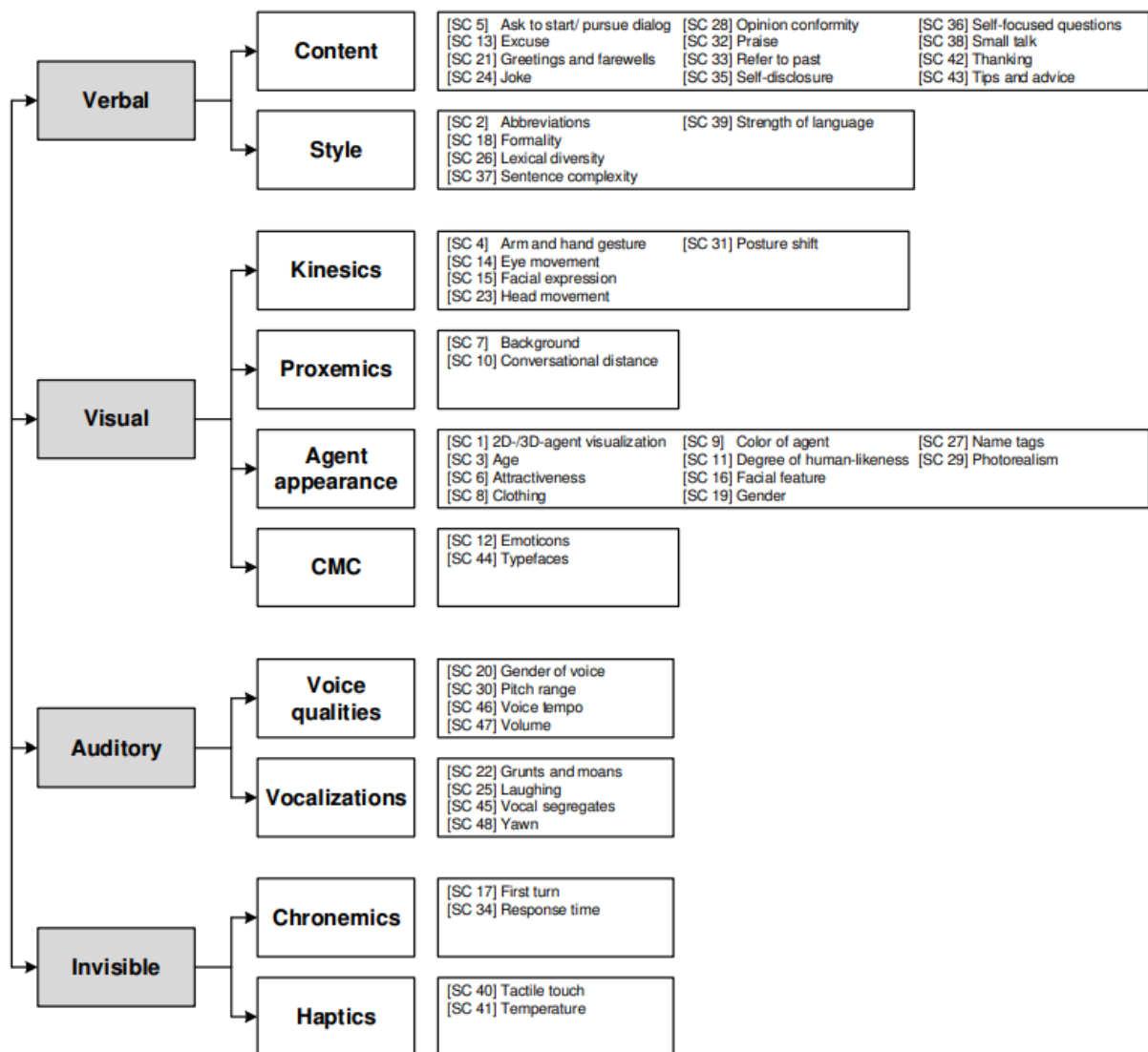
**Figura 8 - Sugestões sociais**

<b>Físicas</b>	· Face, olhos, corpo, movimento
<b>Psicológicas</b>	· Preferências, humor, personalidade, sentimentos, empatia, educação
<b>Linguagem</b>	· Uso interativo da linguagem, linguagem falada, reconhecimento de linguagem
<b>Dinâmica social</b>	· Cooperação, elogio, reciprocidade
<b>Papéis sociais</b>	· Médico, colega de trabalho, adversário, professor, animal de estimação

Fonte: Fogg (2002)

Feine *et al.* (2019) propuseram uma classificação mais abrangente de sugestões sociais em categorias verbais, visuais, auditivas e invisíveis. Para a dimensão verbal, são destacados tanto conteúdo quanto o estilo da linguagem adotada. Como exemplos de ações temos a iniciativa de começar um diálogo, pedir desculpas, agradecer, ou contar piadas. Na forma de se comunicar, é destacada a escolha de palavras utilizadas, com o grau de formalidade adotado, complexidade das sentenças e diversidade léxica.

Figura 9 - Taxonomia de sugestões sociais



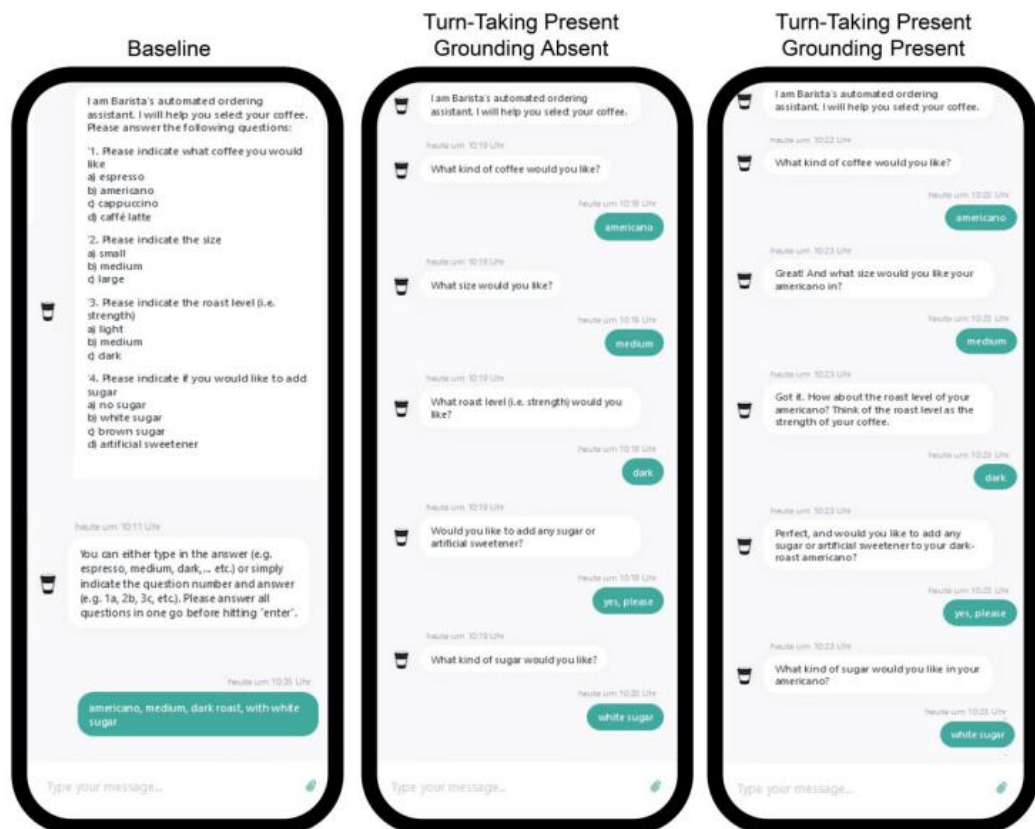
Fonte: Feine *et al.* (2019)

Go & Sundar (2019) realizaram um estudo empírico no qual foram testadas combinações de sugestões visuais, de identidade e conversacionais. O trabalho

concluiu que, na ausência de sugestões visuais, o nível de interatividade das mensagens pode compensar a natureza impessoal do *chatbot*. A interatividade compreende a referência a mensagens trocadas anteriormente de maneira responsiva, trazendo a percepção de um diálogo real e, como consequência, a identificação de maior grau de similaridade entre usuário e agente .

O uso avançado da linguagem, propiciado pelos grandes modelos de linguagem confere uma percepção de autenticidade ainda mais latente aos agentes. demonstram que a aplicação de *chatbot* com propriedades conversacionais semelhantes às humanas promove uma maior aceitação de marcas que utilizam esse meio de comunicação com os consumidores. Dentre as características destacadas estão a alternância de turnos, responsável por garantir que as contribuições ocorrem de forma recíproca entre as partes envolvidas, e iniciação do turno, na qual cada parte controla suas contribuições. Além disso, ao demonstrar partilhar de uma base comum, através de referências ao histórico de interações, os agentes geram reações mais favoráveis.

Figura 10 - Incorporação verbal



Fonte: Bergner *et al.* (2023)

A antropomorfização impacta a forma com que nos relacionamos com a tecnologia. Com o intuito de gerar uma resposta empática, os agentes inteligentes podem ser dotados de características como proximidade, expressões faciais e corporais, e o contexto situacional, capazes de provocar reações emocionais nos usuários (Paiva *et al.*, 2004). Com o advento da inteligência artificial generativa e de agentes inteligentes que mimetizam comportamentos humanos, é necessário ponderar como os usuários interpretam estas sugestões e quais as consequências das intervenções no seu processo decisório (Paiva *et al.*, 2021).

### 2.3.2. Padrões enganosos

A influência exercida por agentes inteligentes levanta debates éticos sobre os impactos das tecnologias digitais. Klenk (2021) define como manipulativas as ações que utilizam formas de persuadir o usuário a expressar determinada conduta. Apesar da tese de que a manipulação está associada à influência social oculta, a preocupação principal está na ausência de divulgação das razões pelas quais as atitudes são incentivadas. Caso o usuário não possua acesso aos motivos que justificam a ação, sua autonomia é comprometida.

Nesse contexto, Gray *et al.* (2018) destacam a necessidade de uma discussão crítica do papel dos *Deceptive patterns* ou padrões enganosos no *design*, práticas que exploram conhecimentos da psicologia para criar funcionalidades que condicionam os usuários a escolhas que favorecem outros *stakeholders*, mesmo quando não alinhadas à estrutura de interesses e valores pessoais. Dentre as possíveis aplicações citadas estão as interferências de interface responsáveis por manipulações estéticas que evocam emoções como estratégia de persuasão.

O experimento realizado por Alberts *et al.* (2024) reforça os riscos de utilizar sugestões sociais de forma inadequada, apontando um potencial efeito contrário ao desejado pelos desenvolvedores:

Em vez de “automaticamente” tratar uma interface social como senciente ou agradável, as pessoas podem, pelo contrário, tornarem-se imediatamente

desconfiadas ao suspeitar que alguém está tentando “conseguir algo” delas. Uma diferença crucial em relação às pessoas como atores sociais é que esperamos que as pessoas ajam em prol de suas próprias necessidades e desejos, enquanto, no caso de uma plataforma, existe a consciência de que ela é um artefato projetado para atender a interesses externos (Alberts *et al.*, 2024).

Tamanha habilidade demonstrada pelos agentes conversacionais levanta questões éticas sobre os possíveis impactos da sua utilização na sociedade, com riscos referentes à privacidade e à ausência de pluralidade, prejudicando usuários pertencentes a grupos socialmente minorizados. Ruane *et al.* (2019) defendem um *design* responsável dessa tecnologia, o que demanda tanto um entendimento aprofundado das características dos grupos de usuários atingidos, quanto a compreensão de seus interesses e contexto de aplicação. Na seção a seguir, levantaremos as principais preocupações inerentes a essa abordagem.

#### 2.4. IA RESPONSÁVEL

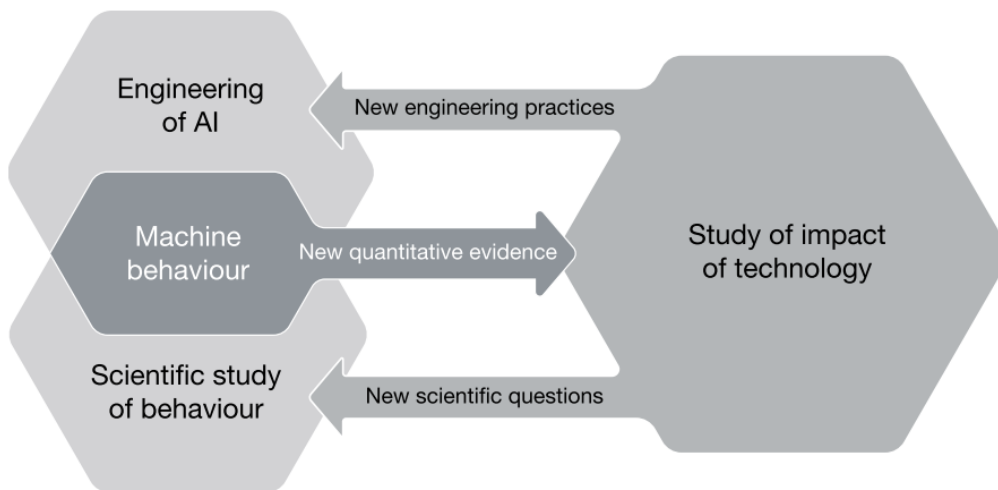
Os agentes de IA têm o potencial de aprimorar o bem-estar humano de várias maneiras e têm demonstrado sucesso em *benchmarks* que avaliam o desempenho em uma série de tarefas. No entanto, pesquisadores do comportamento das máquinas estão interessados em um conjunto mais abrangente de indicadores, investigando como os algoritmos agem em diferentes cenários e se suas interações com humanos afetam as relações sociais, políticas e econômicas (Rahwan *et al.*, 2022).

O desenvolvimento comportamental dos agentes pode ser atribuído a escolhas de engenharia ou design. A seleção das bases de dados de treinamento dos modelos, por exemplo, pode influenciar consideravelmente o comportamento do algoritmo. As máquinas ainda podem adquirir comportamentos por meio de sua própria experiência, como no caso de agentes de aprendizado por reforço. Além disso, o ambiente humano cria forças seletivas que fazem comportamentos bem-sucedidos serem copiados por outros desenvolvedores (Rahwan *et al.*, 2022).

Na implantação de sistemas de IA, os desenvolvedores devem ser guiados pela identificação das relações entre as decisões de projeto e desafios sociotécnicos.

Nesse processo, todas as partes interessadas precisam ser ouvidas, com o objetivo de proporcionar maior segurança às soluções construídas. Uma abordagem democrática, considerando inclusive posicionamentos contrários, enriquece a governança dos ciclos de desenvolvimento por meio de práticas dinâmicas (Dobbe; Krendl Gilbert; Mintz, 2021).

**Figura 11** - Comportamento de máquina



Fonte: Rahwan *et al.* (2022).

Ao tratar dos riscos dos modelos de linguagem, Weidinger et al (2021) destacam as implicações éticas e sociais dessa tecnologia, agrupando-as em seis categorias:

- **Discriminação, exclusão e toxidade:** riscos associados à reprodução de estereótipos ligados a vieses presentes nos dados de treinamento com tendências injustas, tóxicas e opressivas
- **Privacidade:** envolvem violações de privacidade e riscos de segurança, através do vazamento ou inferência de dados confidenciais.
- **Desinformação:** está associada à atribuição de altas probabilidades a informações falsas e ao incentivo de ações não éticas, que podem causar danos, especialmente em áreas sensíveis como o direito e a medicina.
- **Uso Malicioso:** ocorre quando há atuação intencional para disseminação de informações falsas, facilitação de fraudes e crimes, como por exemplo códigos para ataques cibernéticos.

- Interação humano-computador: a utilização dos agentes conversacionais, principalmente influenciada pelo antropomorfismo, promove um excesso de confiança que leva ao compartilhamento de dados e promoção de estereótipos.
- Automação, acesso e danos ambientais: integram as desigualdades na distribuição dos riscos e benefícios da aplicação dos modelos de linguagem entre diferentes grupos sociais, além dos custos ambientais decorrentes da operacionalização dos modelos.

Com uma análise multidisciplinar, Lee *et al.* (2022) abordam a ética em interfaces conversacionais considerando como principais dimensões a divulgação de informações pessoais, a transparência e a vulnerabilidade. As interações com sistemas baseados em padrões de comunicação humanos estimulam o compartilhamento de informações de foro íntimo ao longo do tempo. Diante disso, surge a preocupação quanto à transparência dos processos de coleta e processamento dos dados. Por fim, os autores destacam o papel da diversidade, tanto no sentido de incluir grupos de usuários marginalizados quanto com o objetivo de lidar com usuários que enfrentam algum estado de vulnerabilidade.

Endereçando esses pontos, será possível evitar a reincidência de problemas como os do desastrosos caso da Tay, *chatbot* experimental da Microsoft que foi retirado do ar após dezesseis horas de interação com usuários do Twitter. Tay aprendeu e reproduziu conteúdo em linguagem ofensiva, nazista, antissemita e racista conforme interagia na rede social. O episódio gerou discussões acerca de como esses agentes são percebidos. Inicialmente, o público atribui personalidade e emoção aos chatbots, até que os primeiros erros sejam detectados. A partir desse ponto, outros atores presentes no contexto passam a ser responsabilizados. Se por um lado houve a culpabilização de grupos de usuários que influenciaram o aprendizado de conteúdo que ia desde teorias conspiratórias até posicionamentos supremacistas, por outro os desenvolvedores foram questionados sobre quais medidas poderiam ter sido tomadas ainda em fase de projeto para evitar um comportamento tão suscetível a interferências externas (Neff e Nagy, 2016).

Figura 12 - Tay (Microsoft)



Fonte: Twitter (2006).

O processamento de linguagem natural, incorpora preconceitos históricos e culturais, associando, por exemplo: mulheres e afazeres domésticos, negros e crime, imigrantes e perigo (Bolukbasi *et al.*, 2016; Caliskan-Islam; Bryson; Narayanan, 2017; Durrheim *et al.*, 2023; Garg *et al.*, 2018; Manzini *et al.*, 2019). A ausência de representatividade nos dados de treinamento, nos algoritmos e nos times de desenvolvimento resulta em decisões que amplificam desigualdades já existentes, carregando estereótipos de gênero, étnicos e sociais (Buolamwini; Gebru, 2018; Eubanks, 2018; O'neil, 2016).

A mitigação dos riscos avaliados inclui a aplicação de técnicas capazes de trazer maior segurança às soluções. Dentre elas, está o aprendizado por reforço com feedback humano (RLHF – *Reinforcement Learning from Human Feedback*), que utiliza análises humanas na estrutura de recompensas dos modelos e ajusta seus parâmetros para geração de resultados menos danosos. Outra abordagem consiste no *Red teaming*, na qual grupos multidisciplinares são formados para identificar vulnerabilidades nos sistemas de IA de forma holística. Além disso, métodos de avaliação da IA, em particular dos agentes conversacionais, têm como obstáculo a interpretabilidade, campo que ainda carece de desenvolvimento (Gabriel et al, 2024).

Os aspectos éticos da influência da tecnologia no comportamento humano tornam o estudo deste tipo de relacionamento de extrema relevância para a sociedade. É preciso cautela ao avaliar seus efeitos, pois da mesma maneira que os objetivos por trás da persuasão podem estar associados apenas a questões comerciais, também há outra perspectiva que prioriza o desenvolvimento de tecnologias capazes de promover o bem-estar individual ou coletivo, como ferramentas assistivas ou voltadas ao aprendizado.

### **3. CONFIANÇA PERCEBIDA**

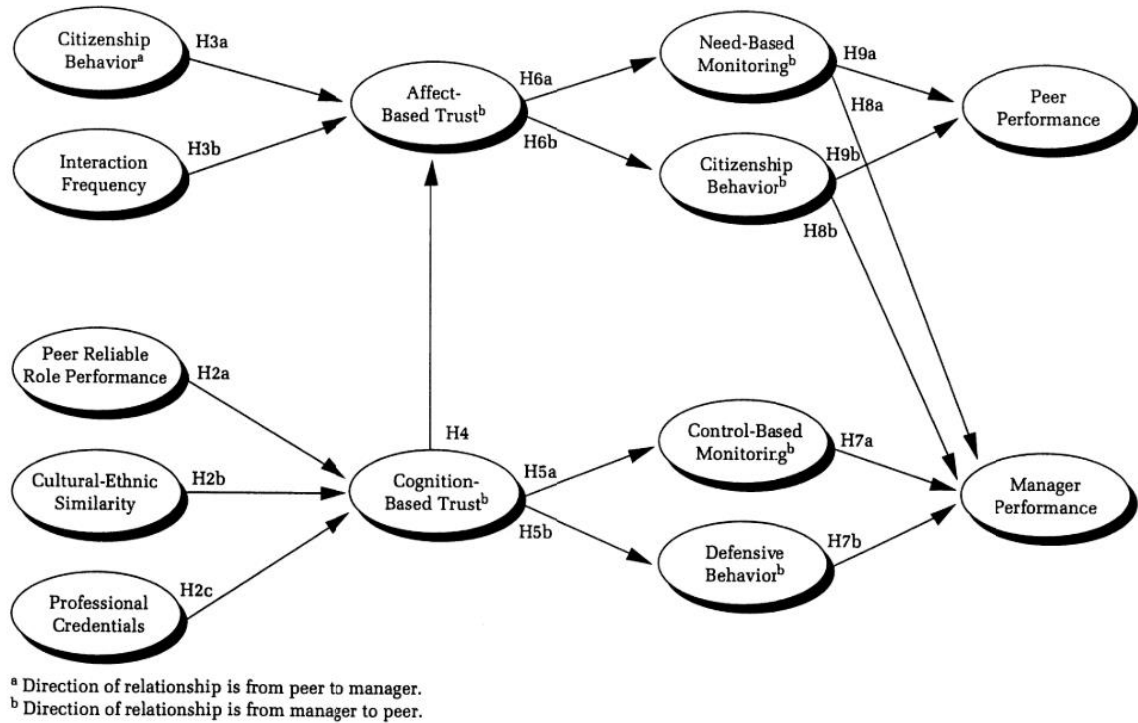
A confiança pode ser definida como “um estado psicológico que compreende a intenção de aceitar a vulnerabilidade com base em expectativas positivas das intenções ou comportamento do outro” (Rousseau *et al.*, 1998, p. 395). Essa visão multidisciplinar, que integra teorias desenvolvidas nas áreas da economia, psicologia e sociologia, apresenta o risco e a interdependência como condições necessárias para a atribuição de confiança. Devido à incerteza das intenções do outro, é necessário que ocorra uma avaliação que leva à assunção ou não do risco. A interdependência, por sua vez, pressupõe que os interesses de uma parte só serão atingidos com a participação da outra.

#### **3.1. PROCESSO DE FORMAÇÃO DA CONFIANÇA**

Ao longo do tempo, pesquisadores expandiram a abordagem dos modelos de confiança seminais propostos por Muir, (1987) e Mayer *et al.* (1995), gerando categorizações para confiança entre humanos e entre humanos e máquinas, que apesar de distintas, carregam similaridades.

Mcallister (1995) propôs um modelo para as relações interpessoais de confiança estruturado em dois fatores: confiança baseada no afeto, que remete aos laços emocionais, e confiança baseada na cognição, associada às avaliações de competência e responsabilidade.

**Figura 13** - Modelo teórico da confiança interpessoal nas Organizações



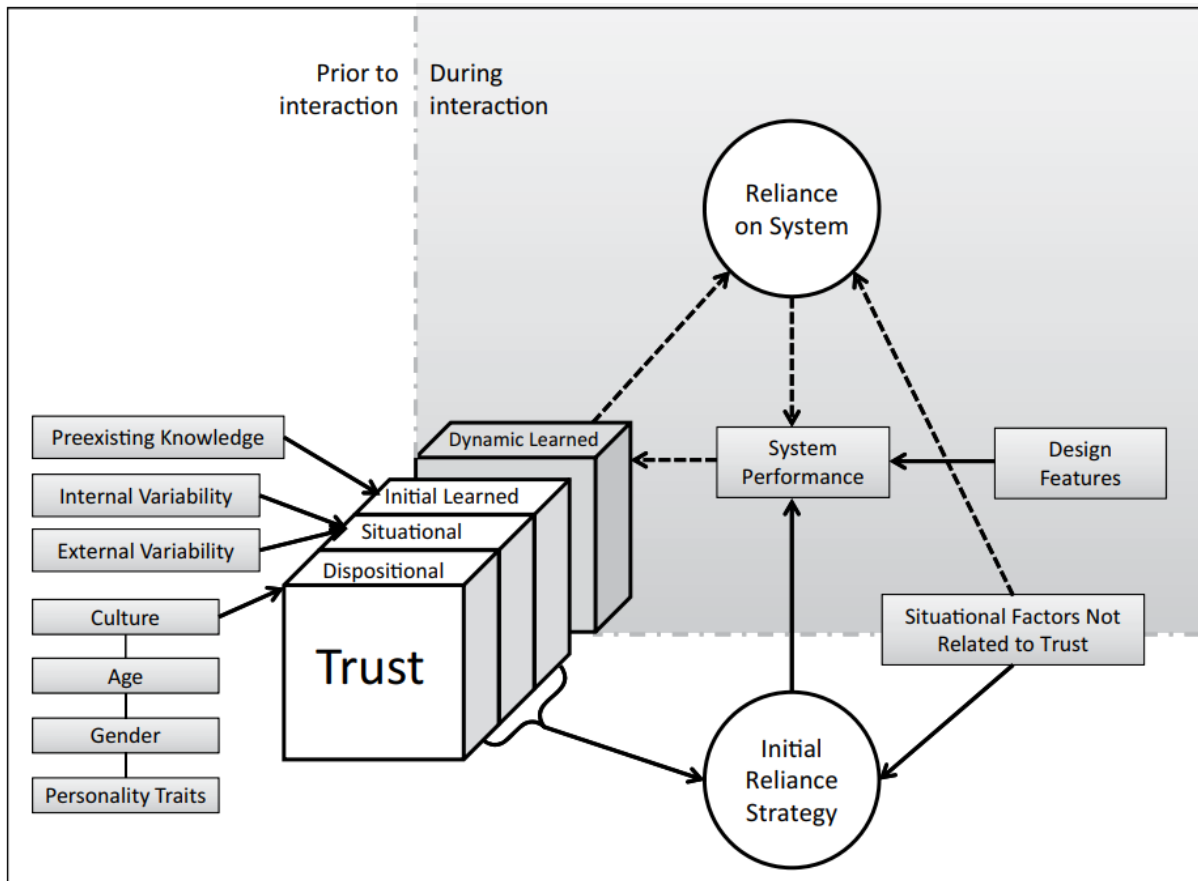
Fonte: Mcallister (1995)

Na análise da interação humano-robô, Hancock *et al.* (2011) assumem uma classificação focada nas características relacionadas ao humano, ao robô e ao ambiente. Os fatores referentes ao robô, sinalizados como os mais relevantes, são agrupados em dois blocos: itens associados a desempenho e itens baseados em atributos, dentre os quais se destacam a personalidade e o antropomorfismo. À época do estudo, os autores já previam uma crescente influência da atribuição de intencionalidade na confiança em sistemas tecnológicos.

Através de estudos empíricos com o uso de *design fiction*, Gulati *et al.* (2019) identificaram os atributos benevolência, competência e risco percebido como significativos para a escala de confiança humano-computador. Analogamente, o modelo proposto por Hoff e Bashir (2015) classifica a variabilidade da confiança na automação em disposicional, situacional e aprendida. A confiança disposicional considera que as avaliações apresentam variações entre grupos demográficos, considerando a interferência de características como cultura, idade, gênero e personalidade. A dimensão situacional, por sua vez, é composta tanto pela variabilidade externa, ligada a fatores ambientais, quanto pela variabilidade interna,

que contempla a autoconfiança do usuário. Por fim, a confiança aprendida é construída a partir das percepções das interações anteriores e das características das interfaces, que também englobam o antropomorfismo.

**Figura 14** - Fatores com impacto na confiança



Fonte: Hoff e Bashir (2015).

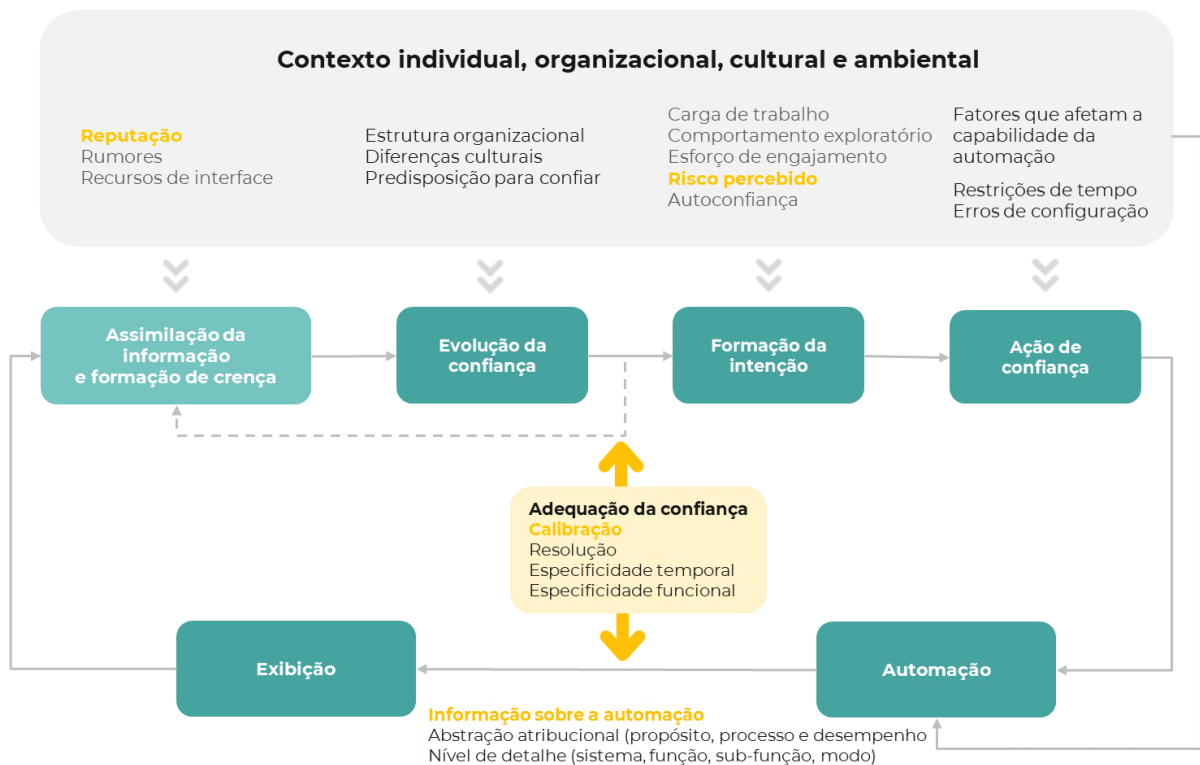
Quanto ao processo de avaliação da tecnologia, contexto, características da automação e processos cognitivos influenciam a assertividade da confiança atribuída. Lee e See (2004) vai além da descrição da dinâmica da confiança, abordando também em que grau esse julgamento está próximo da capacidade da automação, sendo possível identificar os cenários de confiança excessiva e desconfiança.

Conforme ilustrado na figura 15, o processo de confiança passa por algumas etapas. A primeira delas é a de formação da intenção de uso, que envolve majoritariamente informações preliminares e as primeiras impressões de quais são os sinais que aquele agente emite. O momento crítico dessa relação é a violação da confiança, quando uma expectativa de desempenho ou de comportamento é de

alguma forma quebrada. Depois desse ruído, pode haver uma restauração da confiança, que ocorre quando essa atribuição da credibilidade passa por um processo de recuperação e o agente volta a ser considerado como a capaz de seguir com a interação.

O processo de atribuição de confiança considera o contexto, seja individual ou organizacional cultural ou ambiental. Nesse aspecto, é importante ressaltar as informações preliminares que se tem sobre o agente com qual haverá a interação. A reputação os rumores assim como os recursos de interface são considerados. Além disso, características voltadas a estrutura organizacional assim como diferenças culturais e a própria predisposição adotar uma postura mais favorável a confiança interfere nesse processo.

**Figura 15 - Processo de formação da confiança**



Fonte: Lee e See (2004), adaptado

Os primeiros itens citados interferem na formação na crença, enquanto os demais contribuem para a sua evolução. Na etapa seguinte, ocorre a formação da intenção, cujo desenvolvimento está associado a cargas de trabalho bem como a um

comportamento exploratório por parte do usuário, que por meio do esforço necessário para engajamento também identifica o risco associado à interação. Para concretização da confiança, são considerados fatores como a capacidade da automação e as restrições existentes. Por fim, para assegurar a adequação da confiança é preciso considerar o processo de calibração.

Há uma corrente que defende a existência de uma grande diferença entre como humanos e máquinas são julgados. Ainda que haja uma complexidade envolvendo todos os atores, o contexto de decisão e a avaliação de moralidade aplicada, após uma série de experimentos com diferentes cenários referentes à culpabilização de máquinas e humanos, Hidalgo *et al.* (2021) concluíram que “as pessoas julgam humanos com base em suas intenções e máquinas com base em seus resultados”. As observações dos autores vão além do julgamento das intenções. Os estudos apresentados sinalizam que as máquinas são julgadas em um nível correspondente ao dano percebido como resultado da ação. Por outro lado, o julgamento do humano é modulado pela intenção percebida.

Uma visão contrastante e que vem ganhando maior força nos últimos anos, considera que, mediante uma violação da confiança advinda de erros na interação com o humano, a uma retratação por parte da máquina pode atuar na manutenção da intenção de uso através de uma explicação do motivo pelo qual a falha ocorreu. No experimento conduzido por Kox *et al.* (2021), apesar do nível de confiança ter sofrido uma queda após a violação a confiança, pôde ser recuperada com a inclusão de um pedido de desculpas. Essa divergência de comportamento pode ser explicada pelas novas características dos agentes inteligentes e as reações sociais por eles provocadas.

À medida que a tecnologia de IA amadurece, os agentes se tornarão mais sociais e serão mais frequentemente utilizados em papéis sociais. Portanto, parece provável que as pessoas passem a tratar agentes inteligentes como atores sociais, aplicando mais prontamente as mesmas regras, o que pode potencialmente desencadear vieses e heurísticas indesejáveis. O desafio consiste em incorporar habilidades sociais de maneira que apoiem a colaboração entre humanos e agentes, sem induzir os usuários a erro (Kox *et al.* 2021).

Tomando como referência estudos de sistemas de apoio à decisão, uma das principais causas do viés da automação ou da complacência com os erros da IA é a calibração inadequada entre confiança e segurança. Essa relação reflete o estilo

cognitivo do usuário, que pode apresentar uma maior predisposição em aceitar os resultados fornecidos pela tecnologia, com base na sua experiência em interações anteriores. Além disso, aspectos contextuais da complexidade da tarefa, carga de trabalho ou restrição de tempo colaboram com a predominância de heurísticas no processo de avaliação. Para mitigar essa imprecisão, são recomendadas algumas boas práticas no projeto dos sistemas, como a elevação do grau de responsabilização dos usuários e apresentação de informações sobre o a acurácia do modelo, que a rigor compõem o estilo de comunicação praticados (Goddard; Roudsari; Wyatt, 2012).

A fim de investigar em que termos acontece esse julgamento, nas seções seguintes mergulharemos nas métricas e nos mecanismos da cognição social, compreendendo quais são os fatores que interferem nos processos de percepção e a decisão no contexto social.

### 3.2. MÉTRICAS DE AVALIAÇÃO

O comportamento social opera de forma inconsciente ou implícita, sofrendo influência de experiências passadas sem que haja percepção do ator (Greenwald; Banaji, 1995). Por esse motivo, os estudos empíricos realizados no âmbito da cognição social priorizam a utilização de métricas implícitas como meio de avaliar comportamentos inconscientes.

Grande parte da cognição social ocorre de forma implícita. Essa conclusão baseia-se na reinterpretação de diversos achados que destacam a importância do funcionamento implícito das atitudes, especialmente a atitude de autoestima, bem como em evidências já existentes e novas sobre o funcionamento implícito dos estereótipos. Ao integrar essa concepção do modo implícito ao conhecimento existente sobre o modo explícito de operação dos construtos básicos da psicologia social, amplia-se significativamente o alcance desses construtos (Greenwald; Banaji, 1995, p. 20, tradução nossa)

As técnicas experimentais para obtenção de dados implícitos conseguem captar atitudes, crenças ou preferências sem que haja a necessidade de que os indivíduos reflitam e respondam diretamente sobre esses itens. Assim, é possível acessar respostas automáticas através de testes de associação implícita, avaliação de tempos de reação e resultados comportamentais. As medidas explícitas, por outro

lado, incluem pesquisas, questionários e entrevistas em que os participantes fornecem diretamente suas opiniões ou sentimentos sobre o tópico abordado (Fazio, 2003).

Kohn *et al.* (2021) agrupam as medidas de confiança na automação em três categorias: autoavaliações, índices comportamentais e psicológicos. Ao avaliar as medidas utilizadas nos estudos sobre confiança, os autores chamam a atenção para emprego de medidas de autorrelato, recomendando cautela quanto à sua utilização, a fim de garantir a validade das escalas empregadas e a adequação às tarefas e aos objetivos de avaliação. Devido às fragilidades apontadas, esse tipo de indicador é muito simplista, não sendo capaz de capturar o caráter dinâmico e contextual da confiança. Como consequência, há um descompasso entre o comportamento de confiança reportado e o real.

Ainda segundo Kohn *et al.* (2021), a comparabilidade entre os estudos frequentemente é prejudicada pelo uso de medidas personalizadas. Na ausência de uma base teórica às métricas adotadas pode provocar uma insensibilidade a manipulações ou um foco inadequado a aspectos irrelevantes. Diante deste panorama, é possível identificar oportunidades de contribuição por meio do direcionamento a componentes específicos da confiança, seguindo por exemplo aqueles descritos no modelo de Mayer (1995), permitindo prever, explicar e contextualizar os resultados em um arcabouço robusto.

### 3.3. ASPECTOS COGNITIVOS E SOCIAIS

Os diferentes processos cognitivos pelos quais as pessoas entendem e dão sentido ao mundo, contemplando percepção, atenção, memória e planejamento de ações, exercem um importante papel no contexto das interações sociais. Por estar associada aos métodos pelos quais as pessoas compreendem e avaliam informações, a cognição social exerce influência na tomada de decisão. Na avaliação das motivações, objetivos e sentimentos dos outros, os fatores emocionais têm o potencial de facilitar a cooperação. Assim, além de munir o decisor de informações sobre o contexto social no qual ele está inserido, a cognição social guia seu comportamento

para que seus interesses individuais e de grupo sejam atendidos (Frith, 2008; Frith; Singer, 2008).

Mesmo em cenários competitivos, como no caso dos jogos econômicos, as intenções percebidas influenciam a estrutura de preferências do decisor. Considerando que tanto a percepção quanto a resposta a sugestões sociais são moldadas por normas sociais, as pessoas tomam decisões de acordo com as expectativas da sociedade sobre elas, sendo propensas a recompensar as boas e punir as más intenções. O comportamento recíproco, portanto, estimula o altruísmo, em detrimento dos ganhos pessoais, contrariando a teoria econômica clássica (Fehr; Schmidt, 1999).

Assim, é prudente desmistificar a tradicional oposição cartesiana entre razão e emoção ou corpo e mente. Por meio de uma abordagem holística alicerçada na neurociência, Damásio (1994), classifica as emoções como indispensáveis para a tomada de decisão. Precisamente, as estruturas cerebrais responsáveis pelo raciocínio, com destaque para o âmbito pessoal e social também estão associadas às emoções, aos sentimentos e ao processamento de sinais do corpo.

As decisões instintivas são o principal exemplo de como evoluímos para processar rapidamente sinais ambientais e agir assertivamente, sobretudo em situações em que há um risco à sobrevivência. Devido às limitações na capacidade de processamento racional de estratégias de otimização, as decisões humanas recorrem a heurísticas, atalhos mentais que levam a soluções plausíveis. Como todas as opções não são conhecidas, é compreensível que sejam analisadas as que já estão disponíveis e, dentre elas, qual é capaz de gerar maior satisfação. Este é o significado do termo *satisficing*, uma estratégia que simplifica a tomada de decisão com a seleção de soluções suficientemente boas (Simon, 1972; Simon, 1990).

Essas formas de simplificação desenvolvidas pelo cérebro humano para lidar com incertezas envolvem heurísticas como representatividade, disponibilidade e ancoragem. Ao avaliar a probabilidade de um evento tendo estereótipos como referência, a representatividade superestima a previsibilidade e desconsidera a real probabilidade de ocorrência. Analogamente, a disponibilidade atribui um peso maior a eventos recuperados facilmente da memória, fazendo com que eles sejam vistos como mais prováveis, mesmo quando estatisticamente são raros. Por fim, a

ancoragem influencia as estimativas através do acesso a dados iniciais que podem conduzir o decisor a julgamentos tendenciosos (Tversky; Kahneman, 1974).

Ao evidenciar como as escolhas humanas de fato acontecem, as teorias normativas, que retratam as decisões como resultado de um criterioso processo de análise utilitarista, perdem força. O espaço aberto é então ocupado pelas teorias descritivas, elaboradas no intuito de relatar as condições e impactos do julgamento a partir de dados empíricos. Essa abordagem, defendida pela Ciência Comportamental, promove uma quebra de paradigma em direção a uma interpretação construtiva da influência da psicologia e demais áreas do conhecimento na avaliação das alternativas de ação (Thaler, 2016).

A abordagem fenomenológica da confiança em robôs, por exemplo, fornece uma perspectiva que vai além do enfoque meramente racional e instrumental, que se baseia na ideia de escolhas deliberadas e calculadas com base na agência e no controle. A confiança é vista, portanto, como algo inerente às relações sociais. À medida que passamos a interagir e a nos relacionar com essas máquinas, desenvolvemos habilidades e compreensões que nos permitem confiar nelas de maneira mais orgânica, sem necessariamente realizarmos uma avaliação racional de seu riscos limitações (Coeckelbergh, 2012).

### 3.4. JULGAMENTO E TOMADA DE DECISÃO

#### 3.4.1. Viés de autoridade

Os estudos sobre a credibilidade ganharam força com a popularização da internet (Fogg e Tseng, 1999). Ao longo do tempo, estudos sobre a disposição em realizar compras online também foram abordados. Com o avanço das redes sociais, este tema tornou-se ainda mais relevante. Hoje, os motores de busca ocupam um papel já consolidado de ferramentas confiáveis com o status de autoridades cognitivas, percepção obtida heurísticas, ao invés de análises aprofundadas. É provável que os padrões de influência e persuasão sejam replicados aos agentes que operam com a Inteligência Artificial Generativa , porém os estudos neste tópico ainda

estão numa fase inicial, visto que o fornecimento de fontes assertivas que sustentem as respostas, por exemplo, ainda é um assunto em evolução (Li; Sinnamon, 2024).

Ferrario *et. al* (2023) reforçam que os sistemas de IA atuais de fato não podem ser classificados como autoridades, visto que não atendem os critérios necessários para tal. Contudo, o viés de autoridade trata justamente das impressões de autoridade, ainda que infundadas. Cialdini (2007) aponta algumas sugestões de evocam a aparência de autoridade nas relações interpessoais através de símbolos como títulos, roupas e pertences. No contexto dos agentes conversacionais, o viés de autoridade se traduz como uma predisposição à confiança, sem uma avaliação crítica do conteúdo.

Apesar do desempenho impressionante demonstrado pelos sistemas de IA em um número crescente de aplicações do mundo real, a capacidade de gerar previsões precisas não é suficiente para atribuir a eles perícia epistêmica ou autoridade. Na verdade, argumentamos que essas formas de superioridade epistêmica exigem uma relação com o entendimento e um conjunto de habilidades que os sistemas de IA não manifestam (Ferrario *et al.*, 2024).

Em um célebre estudo, Hovland e Weiss (1951) evidenciaram que a fonte da informação tem um peso significativo sobre a opinião dos indivíduos. A exposição a um mesmo conteúdo tende a gerar reações diferentes a depender do grau de credibilidade da fonte. Os achados de Lima *et al.* (2024) corroboram essa constatação. No estudo realizado, quando a fonte não era conhecida, as pessoas avaliaram o conteúdo gerado por humanos e por IA de forma similar. Entretanto, quando exibida a fonte, o conteúdo humano foi priorizado.

#### 3.4.2. Efeito Enquadramento

O modelo de decisão racional tem como pressupostos a extensionalidade e a invariância, pois considera que as escolhas são realizadas com uma avaliação ampla do contexto, ponderando os pormenores do cenário existente, além da potencialidade de riscos e oportunidades em um período futuro. Ambos os princípios presumem que a preferência por determinada opção depende apenas do conjunto de alternativas possíveis para realização da escolha, independentemente de como esse conjunto é descrito, ou seja, os julgamentos não são sensíveis a variações na forma com que as

opções são apresentadas (Arrow, 1982; Kahneman, 2003; Tversky ; Kahneman, 1986).

Contudo, Tversky e Kahneman (1986) evidenciaram que a abordagem normativa não é capaz de descrever os comportamentos de escolha em condições reais, nos quais o efeito de enquadramento acontece.

Variações no enquadramento de problemas de decisão produzem violações sistemáticas de invariância e dominância que não podem ser defendidas em bases normativas. É instrutivo examinar dois mecanismos que podem garantir a invariância de preferências: representações canônicas e o uso do valor atuarial esperado (Tversky; Kahneman, 1986).

As violações ocorrem com os efeitos de enquadramento, dado que descrições extensionalmente equivalentes podem levar a preferências diferentes considerando a saliência relativa de aspectos do problema. Uma situação particular de enquadramento acontece quando uma alternativa é apresentada como padrão, sendo fortemente favorecida frente às demais. O efeito *default*, como é conhecido, tem impacto em decisões importantes por explorar a aceitação passiva da formulação apresentada. Como o decisor não formula representações gerais para as variações de descrição, as escolhas sofrem influência das características que são acessadas com maior facilidade (Kahneman, 2003)

A confiança percebida também está sujeita a oscilações, a depender da maneira como a tecnologia é apresentada. Tratando especificamente da dimensão da competência (habilidade), no estudo conduzido por Gonzalez, Amo e Das Smith (2025), os participantes receberam a descrição de uma sistema de IA que seria utilizado por eles no contexto profissional. As condições experimentais possuíam dois níveis de manipulação. No primeiro, a IA era retratada como fraca ou forte, de acordo com nível de *expertise* e treinamento. No segundo, os resultados gerados eram classificados com base na gravidade dos erros apresentados.

As conclusões do estudo indicam que quando enquadrada como forte, a IA aumenta significativamente a confiança inicial. Ainda que na condição de maior competência consiga manter a confiança frente a imprecisões menores, não houve impacto quanto à tolerância a erros maiores. Na seção a seguir, será possível entender com maior profundidade como a percepção e o julgamento dos agentes ocorre para outras dimensões da confiança.

**Quadro 2:** Enquadramento da competência da IA

Enquadramento 1 – IA Fraca	Enquadramento 2 – IA Forte
<p>Você trabalha na empresa TerraCorp e é responsável por um projeto importante. Você precisa de apoio para analisar os dados do projeto. A TerraCorp tem um sistema proprietário de Inteligência Artificial (IA) que foi desenvolvido por um fornecedor líder em tecnologia. Você decide trabalhar com o sistema de IA na análise dos dados para o projeto. Você fornece ao sistema de IA todos os dados e recebe os resultados no dia seguinte. Você apresenta os resultados aos seus chefes em uma reunião. Na reunião, você descobre que os resultados da análise <b>(não têm erros vs. têm pequenos erros vs. têm grandes erros)</b>.</p>	<p>Você trabalha na empresa TerraCorp e é responsável por um projeto importante. Você precisa de apoio para analisar os dados do projeto. A TerraCorp tem um sistema proprietário de Inteligência Artificial (IA) que foi desenvolvido por um fornecedor líder em tecnologia. O sistema de IA foi desenvolvido especificamente para a análise de dados como os que você precisa de ajuda e tem uma alta taxa de acurácia. Você decide trabalhar com o sistema de IA na análise dos dados para o projeto. Você fornece ao sistema de IA todos os dados e recebe os resultados no dia seguinte. Você apresenta os resultados aos seus chefes em uma reunião. Na reunião, você descobre que os resultados da análise <b>(não têm erros vs. têm pequenos erros vs. têm grandes erros)</b>.</p>

Fonte: Gonzalez, Amo e Das Smith (2025)

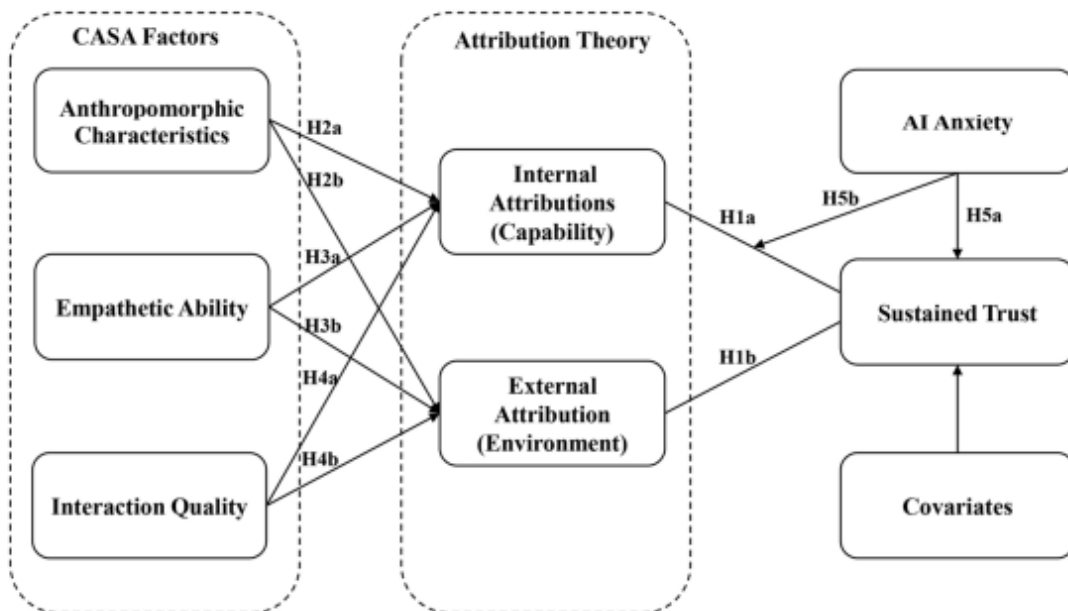
### 3.4.3. Atribuição causal

O processo de percepção dos outros, seus comportamentos e qualidades são básicos nas relações sociais. O estudo do comportamento aparente evidencia que o julgamento das ações observadas sofre o impacto da atribuição de uma origem causadora de determinada conduta, além de ser influenciado por condições do contexto. Apesar do experimento inicial ter sido realizado com a exibição do movimento de figuras geométricas, a hipótese era de que o método poderia ser replicado para a percepção de outras pessoas. A interpretação dos movimentos está associada aos traços de personalidade dos atores, bem como às ações de outros atores ou movimentos anteriores (Heider; Simmel, 1944).

Heider (1956) Define a atribuição como interna ou externa. Na interna, os comportamentos são atribuídos a causas de personalidade. Na atribuição externa, o comportamento é atribuído a questões contextuais. É possível traçar um paralelo com os erros de sistemas automáticos. Quando há uma relação de maior proximidade, a confiança é resiliente mesmo diante do erro.

O modelo proposto por Gu *et al.* (2024) relaciona os fatores característicos do paradigma CASA (Computadores como atores sociais) com a teoria da atribuição, tendo a resiliência da confiança a eventuais erros do agente como resultante dessa interação. A pesquisa empírica realizada com 500 participantes, concluiu que as sugestões sociais apresentadas por *chatbots* de serviços são capazes de manter a confiança dos usuários mesmo em cenários de falhas. Assim, a confiança está positivamente relacionada com a atribuição externa e negativamente relacionada com a atribuição interna. Foram avaliados fatores sociais como características antropomórficas, habilidades empáticas e qualidade da interação, além do fator de ansiedade à IA. As características antropomórficas influenciam a percepção do usuário de modo a associar as falhas dos *chatbots* a fatores externos/ambientais, não impactando a percepção da sua capacidade. O mesmo acontece com a empatia e com a qualidade da interação. Ansiedade na IA é apresentada como moderadora, potencializando os efeitos da atribuição interna na confiança.

**Figura 16** - Modelo da confiança sustentada



Fonte: Gu *et al.* (2024)

#### 3.4.4. Antropomorfismo

O Antropomorfismo, tendência atribuir traços humanos a entidades não humanas, considera, além de características físicas, as emoções associadas a uma agência moral como consciência, intenção, desejo e arrependimento. Esse efeito é reforçado por fatores psicológicos a partir do conhecimento elicitado do agente e da motivação do avaliador, tanto sobre os efeitos esperados da interação com o agente quanto às necessidades sociais. O fenômeno tem implicações importantes para o *design* de tecnologia e o entendimento de comportamentos culturais e educacionais (Epley *et al.*, 2007).

De Visser *et al.* (2016) identificaram que agentes antropomórficos estão associados com maior resiliência da confiança atribuída pelos usuários. Tal relação se torna ainda mais significativa em cenários de maior incerteza. Isso quer dizer que na presença do antropomorfismo há uma expectativa de que as máquinas sejam mais assertivas. Apesar disso, quando são atribuídas características humanas aos agentes, há uma maior permissividade em relação aos erros cometidos por eles. O estudo identificou que aparência e o comportamento dos agentes inteligentes provocam alterações no antropomorfismo percebido, mas que outras dimensões como a agência, intencionalidade e presença física também podem ser utilizadas. A representatividade do agente também é pontuada como um fator que merece atenção características como idade podem promover uma maior confiança assim como idade, gênero, etnia e voz. Para aprofundar a compreensão dos motivos que levam a tamanhos efeitos, utilizando os constructos instância intencional, parassocialidade e presença social.

A instância intencional, termo cunhado por Dennett (1988), parte do pressuposto que temos a necessidade de compreender e prever o comportamento dos agentes com os quais interagimos, sejam eles humanos, animais ou até máquinas. No intuito de dar sentido às complexas relações presentes no meio ambiente, ainda que não haja conhecimento preciso sobre mecanismos causais, lançamos mão da estratégia de atribuir estados mentais como crenças e desejos que motivariam as ações dos sistemas. Tal definição também possui forte base empírica, tendo em vista trabalhos como o desenvolvido por Heider e Simmel (1944) que

evidenciaram o caráter subjetivo da atribuição de significado aos estímulos recebidos do ambiente na percepção e interpretação de “comportamentos aparentes”. O processo de associar intenções, personalidades, e emoções a seres inanimados a partir de inferências do contexto social nos ajudam a entender como os padrões de julgamento são aplicados.

**Figura 17 - Antropomorfismo**

Categoria das variáveis	Principais determinantes psicológicos		
	Conhecimento elicitado do agente	Motivação de efeito	Motivação social
Disposicional	Necessidade de cognição	Necessidade de proximidade Desejo de controle	Solidão crônica
Situacional	Semelhança percebida	Interação antecipada previsibilidade aparente	Desconexão social
Desenvolvimento	Aquisição de teorias alternativas	Alcançar competência	Apego
Cultural	Experiência normas e ideologias	Evitar a incerteza	Individualismo e coletivismo

Fonte: Epley *et al.* (2007), adaptado.

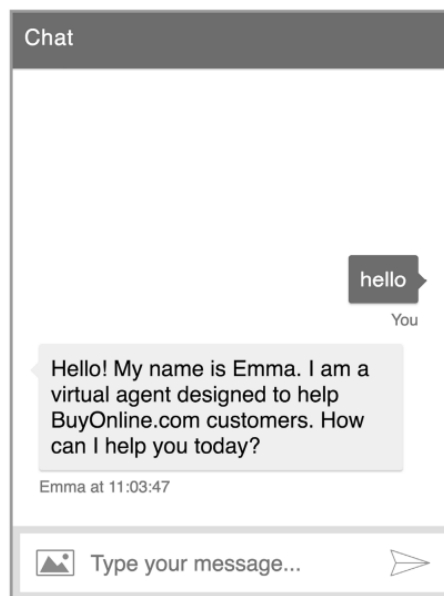
Analogamente, a parassocialidade, conceito aplicado ao relacionamento unilateral entre indivíduos e personalidades da mídia, personagens ou celebridades nos ajuda a compreender a interação dos indivíduos com os agentes conversacionais. Apesar do caráter ilusório, há o desenvolvimento de uma conexão pessoal com a figura midiática, o que gera impactos na percepção de naturalidade e, consequentemente no grau de confiança atribuído a estes agentes. Portanto, a antropomorfização de *chatbots* aumenta a receptividade dos usuários às informações fornecidas, tornando-os menos propensos a realizar a verificação dos resultados. Com a falsa impressão de reciprocidade, a parassocialidade indica como os espaços mediados construídos pela IA generativa interferem nas emoções humanas e se beneficiam disso para alavancar sua utilização (Maeda; Quan-Haase, 2024).

A presença social é definida como a sensação de “estar junto com o outro”, incluindo respostas primitivas a sugestões sociais, simulações de “outras mentes” e

modelos gerados automaticamente da intencionalidade dos outros (Nowak; Biocca, 2003). As características podem ser físicas na forma de avatares ou robôs, que mimetizam expressões faciais e o corpo humano, mas também podem significar atribuição de personalidade de traços é semelhante ao humano, principalmente no discurso, o que foi possibilitado através dos modelos de linguagem e pela inteligência artificial generativa.

Em um experimento realizado com um *chatbot* no contexto de compras online, Araujo (2018) identificou que a utilização de sugestões sociais contribui para a percepção de antropomorfismo pelos usuários, ainda que os agentes não possuam uma representação física. As sugestões sociais impactaram positivamente a conexão emocional dos usuários com a empresa. Contudo, a presença social percebida não sofreu alterações significativas entre as condições de *chatbot* humanizado x *chatbot* mecanizado.

**Figura 18** - Interface de chat antropomórfico

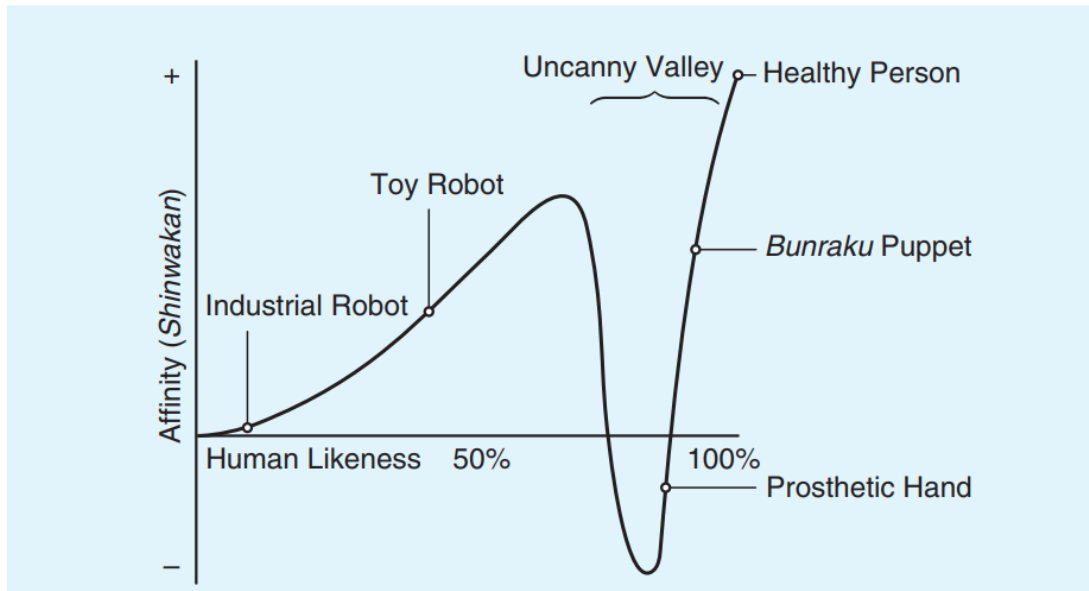


Fonte: Araujo (2018)

É importante destacar, contudo, que a relação dos usuários com agentes dotados de características antropomórficas nem sempre é positiva. No contexto da robótica, a hipótese do “vale misterioso” defende que a afinidade com as características humanas cresce até determinando ponto. Quando há a percepção de

que o agente carece de características esperadas, a afinidade sofre uma profunda queda, tendo como consequência uma sensação de estranheza (Mori *et al.*, 2012).

**Figura 19** - Vale misterioso



Fonte: Mori *et al.* (2012).

### 3.4.5. Modelo de conteúdo de Estereótipos

O julgamento social é permeado por preconceitos. Pioneiro na abordagem do tema, o psicólogo Gordon Allport define que "Preconceito é uma antipatia baseada em generalização falha e inflexível. Pode ser sentido ou expresso. Pode ser direcionado a um grupo ou a um indivíduo desse grupo" (Allport, 1954).

Derivando da teoria de julgamentos sociais de Rosenberg *et al.* (1968), que classificou os traços de personalidade nas dimensões intelectual e social, o Modelo de Conteúdo de Estereótipos estabelece cordialidade e a competência como dimensões centrais da percepção social. Enquanto a cordialidade engloba a sociabilidade e a moralidade, a competência envolve os fatores de agência e habilidade. Assim, a cordialidade percebida é prevista pela relação de interdependência (cooperativa ou competitiva) e a competência percebida pelo status. Tradicionalmente ligados a percepções negativas, os estereótipos também podem gerar reações subjetivamente positivas. Ao analisar os componentes psicológicos do

preconceito, é possível afirmar que as emoções atuam como mediadoras entre cognição e discriminação, gerando diferentes combinações (Cuddy; Fiske; Glick, 2008).

As combinações entre as percepções de competência e cordialidade podem ocorrer em duas direções. Na associação negativa, as dimensões são avaliadas em direções opostas, gerando estereótipos ambivalentes, ou seja, de baixa competência e alta cordialidade ou baixa cordialidade e alta competência. Uma relação positiva, no outro extremo, ocorre quando ambos os eixos convergem para um mesmo padrão de julgamento, podendo haver inclusive a interferência de uma das dimensões na percepção da outra, relação explicada pelo efeito *halo*, detalhado na seção seguinte (Biancardi; Cafaro; Pelachaud, 2017).

No primeiro quadrante (Figura 20), com uma alta cordialidade e baixa competência, temos grupos como crianças e idosos, sobre os quais é evocado o sentimento de pena. Quando a competência é baixa e a cordialidade também é baixa, como por exemplo em grupos sociais a imigrantes e pessoas pobres, é despertado um sentimento de desprezo. No outro extremo, com alta competência e alta cordialidade, temos a admiração. Isso ocorre em situações em que homens brancos por exemplo são tidos como sinônimo de sucesso. Por fim, no quadrante onde há uma alta competência, porém uma baixa cordialidade, está localizada a inveja evocada no julgamento de grupos como asiáticos ou judeus (Fiske *et al*, 2002).

Embora ambas as dimensões sejam fundamentais para a percepção social, os julgamentos cordiais são prioritários, refletindo a importância de avaliar primeiro as intenções antes de determinar a capacidade do outro para concretizar essas intenções. A cordialidade se sobressai não apenas pelo fato de ser julgada antes da competência, mas também por possuir maior peso nas relações afetivas e comportamentais. Avaliar as intenções dos outros, seja para o bem ou para o mal, é mais importante que prever a sua capacidade de ação. Analogamente, a dimensão da moralidade é priorizada nos julgamentos, visto que está associada a um conhecimento acessado com maior facilidade (Cuddy; Fiske; Glick, 2008).

**Figura 20** - Modelo de conteúdo de estereótipo



Fonte: Fiske *et al* (2002)

Muitos dos traços de julgamento social envolvem a moralidade, como por exemplo a percepção de alguém como sincero, prestativo e honesto. Nessa avaliação, as informações negativas são mais decisivas, dado que o comportamento imoral de um indivíduo tende a ser prejudicial para os demais membros do grupo social. Assim, a formação de impressões está relacionada à função de identificar as pessoas das quais se aproximar e aquelas das quais se afastar (Wojciszke; Bazinska; Jaworski, 1998).

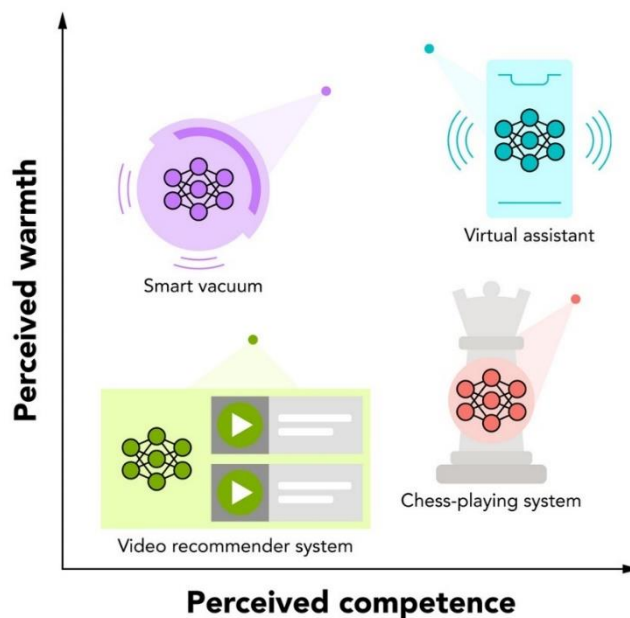
A avaliação moral faz parte, portanto, de um mecanismo de vigilância automática, aprimorado ao longo da evolução humana para identificar ameaças sociais no meio ambiente. Por isso, conforme já comprovado através de estudos empíricos, a resposta a estímulos voltados à competência tende a ser mais lenta. Independentemente da idade, as pessoas são mais sensíveis aos sinais que sugerem riscos quanto à moral e às intenções dos outros (Ybarra *et al.*, 2001)

Apesar de classificar competência e cordialidade como dimensões universais, o modelo de conteúdo de estereótipos sofreu críticas por estar inserido no padrão de sociedade dos EUA. Tendo em vista a multiplicidade de culturas e valores ao redor do mundo, seria difícil encontrar uma uniformidade de avaliações. A fim de suprir e referida limitação, Cuddy *et al.* (2009) conduziram um estudo mais amplo, com

peças de diferentes nacionalidades, e identificaram que o modelo permanecia válido nesse contexto.

O *framework* do SCM permanece válido entre culturas, prevendo como os grupos provavelmente são estereotipados com base em suas relações estruturais com outros grupos na sociedade. Aplicar esse *framework* de forma sistemática e válida em diferentes culturas evidencia diferenças culturais importantes, como a ausência de grupos universalmente favorecidos em culturas coletivistas. Assim, as semelhanças e diferenças culturais ampliam os entendimentos sobre as relações intergrupais, indo além de perspectivas centradas nos EUA (Cuddy *et al.*, 2009).

**Figura 21** - Percepção social de agentes



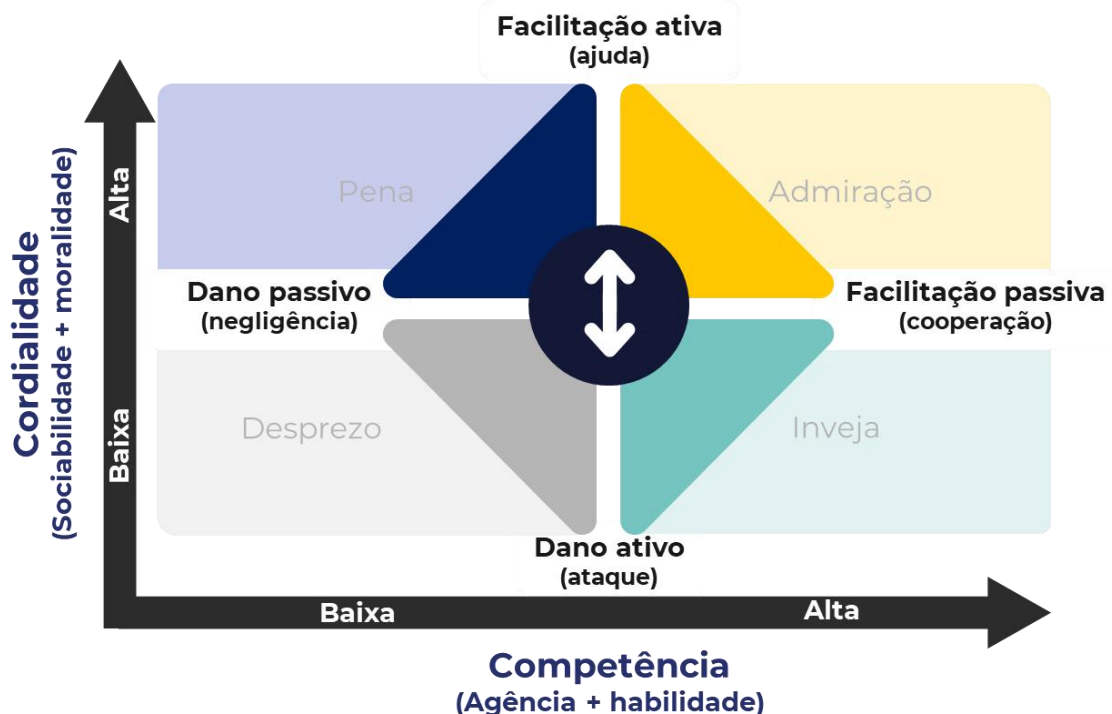
Fonte: McKee *et al.* (2023).

Diante da adoção de características sociais na concepção de novas tecnologias, a forma de avaliação pelo usuário adquire um perfil semelhante ao utilizado para classificar as pessoas. Agentes artificiais projetados com sugestões de alta competência, por exemplo, geralmente evocam reações afetivas, comportamentais, atitudinais e cognitivas, tendo como consequência maior receptividade aos resultados fornecidos durante a interação (Liew; Tan, 2021). Gilad *et al.* (2021) identificaram que a preferência por cordialidade em detrimento de competência também é replicada no contexto da avaliação de sistemas de IA, inclusive nos casos em que o sistema possui um baixo desempenho. Essa percepção

sinaliza que o julgamento da IA que exibe características associadas a humanos se assemelha à dinâmica das relações interpessoais.

Com o surgimento de novos papéis para a tecnologia, McKee *et al.* (2023) trazem uma reflexão sobre o papel na cordialidade e a intencionalidade percebida para atribuição do grau de confiança aos agentes inteligentes. Traçando um paralelo com o modelo de confiança de Mayer(1995), a competência está para a habilidade, assim como a cordialidade está para a benevolência e integridade. Ao analisar a receptividade de humanos a agentes virtuais como colegas de trabalho, Harris-Watson *et al.* (2023) encontraram uma importante relação entre a cordialidade e competência com os componentes de aceitação. Analogamente, no contexto da IA como assistente virtual no processo de compras *on-line*, Kim e Hur (2024) identificaram que, devido à necessidade de interação, fatores como personalização do *chatbot* e antropomorfismo contribuíram significativamente para a percepção de competência e cordialidade pelos consumidores, levando à empatia e aceitação da ferramenta de comunicação.

Figura 22 - Bias Map



Fonte: Cuddy *et al* (2008)

Cada combinação gera um padrão de comportamento (Figura 22). A depender do grau de cordialidade e competência percebida, é gerado o comportamento de

facilitação ou de dano. Para ambos, o processo pode ocorrer de forma ativa ou passiva. Para os casos de alta competência, a facilitação passiva que estimula comportamentos de cooperação. Por outro lado, quando há baixa competência, o dano é passivo, que está associado em situações de negligência. Com foco na alta cordialidade, temos a facilitação ativa, na qual ocorre a ajuda aos membros dos grupos identificados como o mais sociáveis ou morais. Por fim, quando a cordialidade é baixa, corre o dano ativo, que consiste no ataque aos grupos presentes neste quadrante.

## **4. CONFIANÇA APROPRIADA**

### **4.1. CALIBRAÇÃO**

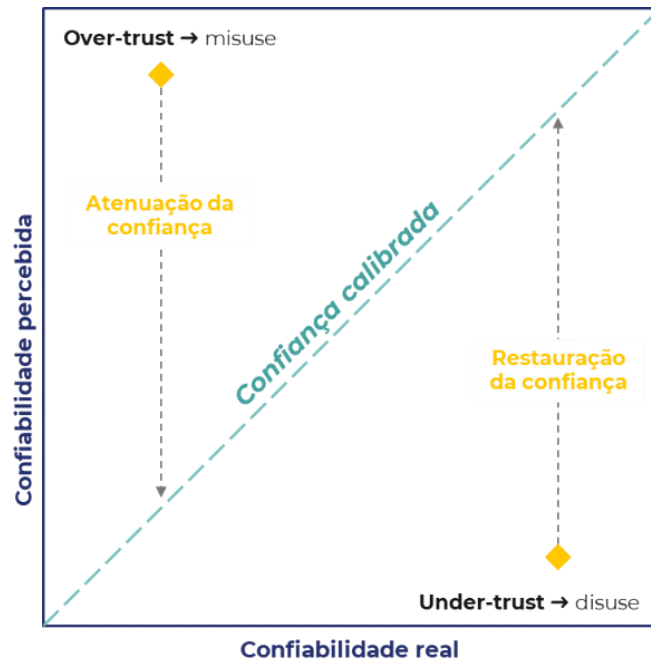
A confiança não se comporta como um atributo estático, sendo o produto de um processo dinâmico, sujeito variações e ajustes. Por este motivo, a calibração merece um destaque especial neste trabalho. Para facilitar o entendimento deste papel, De Visser (2020) apresenta de forma gráfica as distorções existentes entre a confiabilidade percebida e a capacidade real dos agentes. Os autores destacam a execução de ações capazes de atenuar o nível de confiança atribuído nos casos de excesso de confiança ou *misuse* e fatores capazes de restaurar a confiança quando a classificação dos agentes aponta para um nível aquém da capacidade de real, fazendo com que se alcance o nível de calibração da confiança.

A resiliência, por outro lado, está preocupada em entender até que ponto a confiança permanece inabalada, mesmo diante de eventuais sinais que possam levar a um processo de quebra de confiança. Contrariando a tendência de aplicação de estratégias para elevação do nível de confiança atribuído aos agentes, uma abordagem responsável pode até aumentar a percepção de risco na tecnologia, quando aplicável. Ao longo deste capítulo, serão apresentadas algumas alternativas capazes de realizar essa aproximação.

Liao e Sundar (2022) defendem que a comunicação da confiança é um fator bastante relevante, visto que os sistemas apresentam sugestões que evocam ações com impacto direto nessa avaliação. Da mesma forma que pode haver práticas que

influenciem a atribuição de um confiança além da capacidade real da tecnologia, podemos ter casos de um nível de desconfiança excessivo. Há então a necessidade de que os usuários tomem decisões conscientes. Na outra ponta, é preciso que os responsáveis pela concepção e desenvolvimento dos sistemas considerem estes aspectos de maneira ética.

**Figura 23** - Calibração da confiança



Fonte: De Visser (2020)

Características como reputação e o próprio antropomorfismo podem levar a uma avaliação irreal da confiabilidade dos sistemas. Por isso, os mecanismos de ajuste do nível de confiança são necessários. Tomando como referência o início da interação, a calibração exógena envolve a aplicação de intervenções prévias, informando os usuários sobre limitações, capacidade e reputação do sistema. Por outro lado, a calibração endógena traz dados referentes à acurácia dos modelos enquanto as interações estão em curso. Além disso, o ajuste do nível de confiança pode ser realizado de maneira estática ou adaptativa (Wischnewski *et al.*, 2023).

No contexto da colaboração humano – IA, Okamura e Yamada (2020) apresentam um framework no qual são avaliadas a probabilidades de acerto do agente, a expectativa de acerto do agente atribuída pelo humano e a probabilidade de acerto do humano. Na interação com agentes conversacionais, o perigos da confiança excessiva na automação incluem a complacência, a perda de controle, comportamento impulsivo, além da própria confiança descalibrada.

**Tabela 1** - Classificação do tipo de intervenção para calibração da confiança

Momento da intervenção	Intervenção
Antes da interação	Informações prévias sobre o sistema como capacidade, confiabilidade, consistência, desempenho, defeitos, fase de treinamento ou esforço cognitivo.
Durante a interação	Sinalização por meio de alarmes, alertas, exibição de incerteza, estilo de comunicação, explicações, interatividade
Depois da interação	Tipo de <i>feedback</i> ( padrão/personalizado); avaliação de desempenho.

Fonte: Wischnewski *et al.*(2023)

Para modelos de linguagem em particular, o *gap* entre o conhecimento real e o conhecimento percebido pelo usuário é crítico, dada a sua aplicação nas mais diversas áreas. A tendência de superestimação da acurácia dos modelos de linguagem está associada a explicações padronizadas. Assim, as explicações personalizadas conseguem ser mais efetivas, controlando a expressão de incerteza nas respostas, trazendo transparência ao processo (Steyvers *et al.*, 2024).

Os usuários frequentemente superestimam a precisão das respostas dos LLMs, especialmente ao confiar nas explicações padronizadas. Isso foi observado em dois LLMs diferentes, destacando uma tendência preocupante de confiança excessiva nas capacidades dos modelos, o que pode ser prejudicial em contextos em que decisões críticas dependem dessas informações. A principal contribuição deste estudo foi demonstrar a eficácia de explicações personalizadas para reduzir essa lacuna de percepção. Alteramos os prompts usados para gerar explicações baseadas na confiança do modelo, controlando a expressão de incerteza nas respostas. Por exemplo, prompts ajustados variaram de declarações de baixa confiança (“Não tenho certeza de que a resposta seja [B] porque”) a alta confiança (“Tenho certeza de que a resposta é [B] porque”). Essa abordagem ajudou os usuários a calibrar melhor suas avaliações da confiabilidade do modelo e a distinguir respostas corretas de incorretas. Isso reforça a importância da comunicação transparente por parte dos LLMs e sugere que a forma como explicações são apresentadas pode impactar significativamente a percepção dos usuários (Steyvers *et al.*, 2024, tradução nossa).

As sugestões propostas por Dubiel *et al.* (2022) para mitigar os riscos da confiança excessiva nos agentes conversacionais incluem a priorização do controle do usuário frente às funções do agentes, sendo possível, por exemplo, habilitar ou desabilitar recomendações, ajustando o nível de apoio recebido durante a interação. Ao promover reflexões e sugerir alternativas, os agentes estimulam os usuários a identificar as motivações por trás de suas escolhas. Os autores desencorajam a implementação de características antropomórficas para garantir que os usuários tenham expectativas mais realistas quanto às limitações dos modelos.

#### 4.2. ESTRATÉGIAS DE DEBIASING

O desviesamento, ou *debiasing* na literatura em língua inglesa, propõe a aplicação de intervenções que ampliem a compreensão dos mecanismos envolvidos no processo decisório (Fischhoff, 1982). A IA generativa torna ainda mais relevante um ajuste da confiança na avaliação dos resultados. Características como a extensão e a facilidade de geração de conteúdos, os modos de falha associados à sua natureza não determinística e a dificuldade de definir medidas de qualidade objetivas oferecem desafios às dinâmicas de calibração.

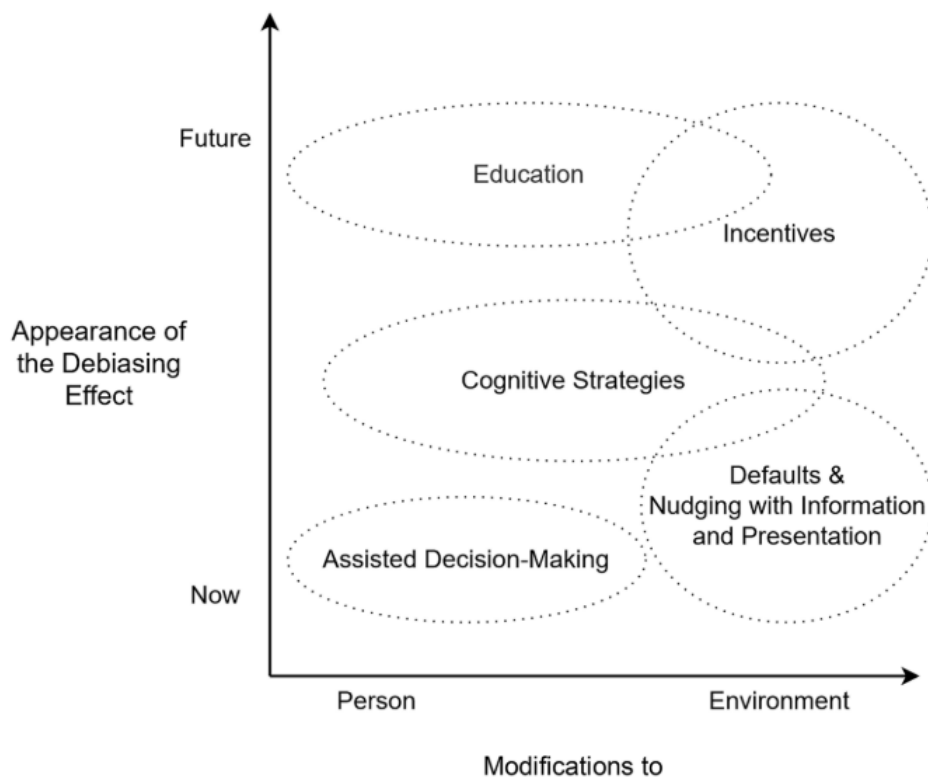
Ji (2024) defende a necessidade de desmistificar o antropomorfismo associado à IA generativa, visto que ele ofusca os reais perigos e impactos sociais dessa tecnologia, potencializando a difusão de desinformação. Portanto, é imprescindível construir uma representação genuína da IA, que transforme a maneira com que ela é percebida e utilizada, sobretudo nos cenários associados ao excesso de confiança.

Embora existam intervenções mais urgentes para modelos generativos, o antropomorfismo continua a influenciar questões críticas relacionadas à IA. Por exemplo, o antropomorfismo agrava a disseminação de desinformação, já que sistemas de IA antropomorfizados tendem a ser percebidos como mais humanos e confiáveis. Bots de IA podem ser facilmente aceitos, o que aumenta a criação de mídias enganosas. Assim, corrigir o antropomorfismo na IA não é apenas sobre construir uma representação genuína desses sistemas, mas também essencial para promover uma compreensão informada sobre a IA. (Ji, 2024, p.7, tradução nossa).

Apesar de diversas quanto ao propósito e aos meios de implementação, as estratégias de *debiasing* são complementares. Para categorizá-las, é possível considerar duas dimensões: o momento em que a intervenção ocorre e o tipo de modificação que ela causa. As ações voltadas à educação demonstram efeitos a longo

prazo no âmbito pessoal, dado que o letramento na IA e em temas como privacidade e proteção de dados promovem a maturidade do usuário à medida que há a construção de conhecimento sobre os tópicos. As abordagens cognitivas, por outro lado, têm como objetivo gerar efeitos imediatos a nível ambiental, já que são aplicadas quando a interação acontece, modificando seu contexto (Leschanowsky; Popp; Peters, 2023).

**Figura 24** - Classificação das estratégias de debiasing



Fonte: Leschanowsky *et al.* (2023).

As primeiras impressões sobre um modelo contribuem para a avaliação de confiança e intenção de uso de um sistema inteligente. Quando as explicações disponibilizadas destacam pontos fortes do sistema, os usuários tendem a desenvolver um maior nível de confiança, desconsiderando os potenciais riscos associados à interação. Por outro lado, quando os pontos fracos são apresentados inicialmente, o usuário tende a descredibilizar o sistema (Nourani *et al.*, 2021).

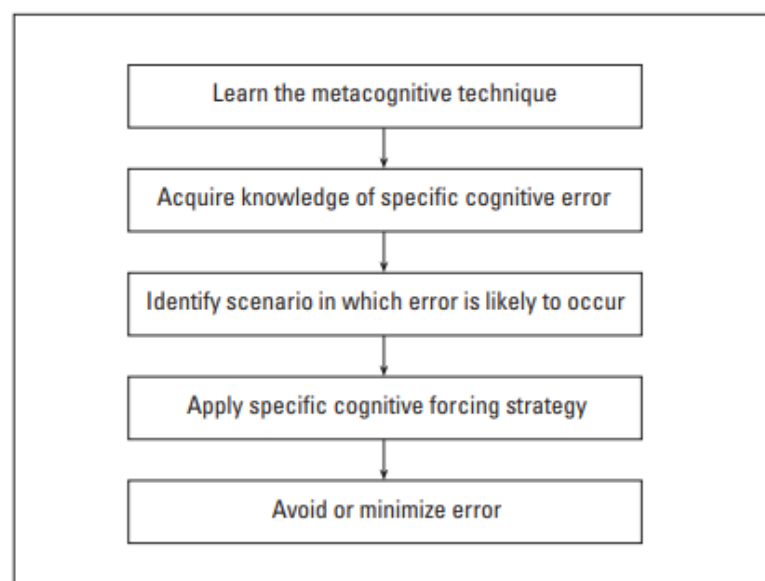
Fundamentada pela teoria dos processos cognitivos duais, a abordagem do esforço cognitivo envolve a aplicação de intervenções que promovem o engajamento analítico dos usuários nas interações. Quando comparada com intervenções comuns

de IA explicável, a adição do esforço cognitivo é capaz de reduzir o excesso de confiança na IA. Contudo, pode haver uma percepção negativa do usuário quanto a seu nível de complexidade. Considerando a natureza probabilística das aplicações baseadas em modelos de linguagem, a apresentação do nível de incerteza inerente ao seu processo de geração de respostas pode contribuir com avaliações de confiança mais calibradas (Buçinca *et al.*, 2021).

Ainda que considere um tipo de aplicação diferente, o conhecimento gerado em trabalhos sobre sistemas de apoio a diagnóstico, também trata do excesso de confiança e pode fornecer *insights* sobre meios que aproximem a avaliação dos sistemas de sua capacidade real. Especificamente para atuações voltadas ao local de trabalho, há histórico da utilização de checklists, estratégias de esforço cognitivo, reflexão guiada, instruções para avaliação analítica (Lambe *et al.*, 2016).

A adoção consciente e deliberada de uma estratégia para otimizar a tomada de decisão evita a incidência de erros associados a heurísticas e vieses. Tomando como referência a tomada de decisão no contexto clínico, os médicos podem tomar decisões mais assertivas implementando táticas de esforço cognitivo, que incluem a avaliação de diferentes alternativas, a fim de garantir uma avaliação cuidadosa antes de tirar conclusões de forma precipitada. Na figura abaixo, são listadas as etapas sugeridas para implementação de uma estratégia de esforço cognitivo (Croskerry, 2003).

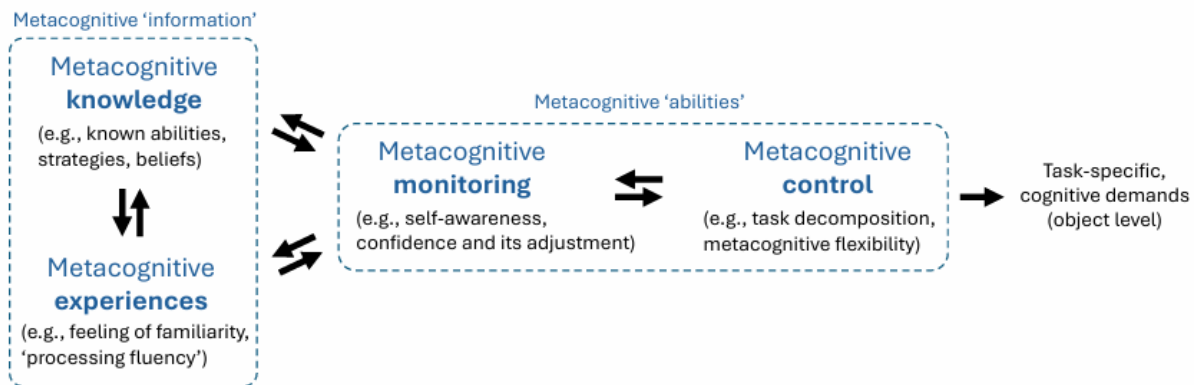
**Figura 25** - Etapas para estratégia de esforço cognitivo



Fonte: Croskerry (2003).

Além disso, é possível aplicar intervenções que apoiem a metacognição do usuário, ou seja, a habilidade de entender de forma consciente seu processo de decisão. Conforme exibido na figura a seguir, o conhecimento e as experiências metacognitivas se retroalimentam, de forma que a cada experiência o conhecimento vigente é recuperado e novos aprendizados são gerados. Ambos se relacionam com o monitoramento metacognitivo, que promove a autoconsciência e o ajuste da confiança. Com base nesse acompanhamento que o controle metacognitivo é disparado, atuando sobre as tarefas desempenhadas. Alguns exemplos são a resolução de problemas por etapas ou a exibição de *feedback* para ajuste do modelo mental formado pelo usuário, estimulando o pensamento crítico e contribuindo para uma melhor avaliação da confiabilidade dos resultados obtidos (Tankelevitch *et al.*, 2024).

**Figura 26** - Framework para metacognição



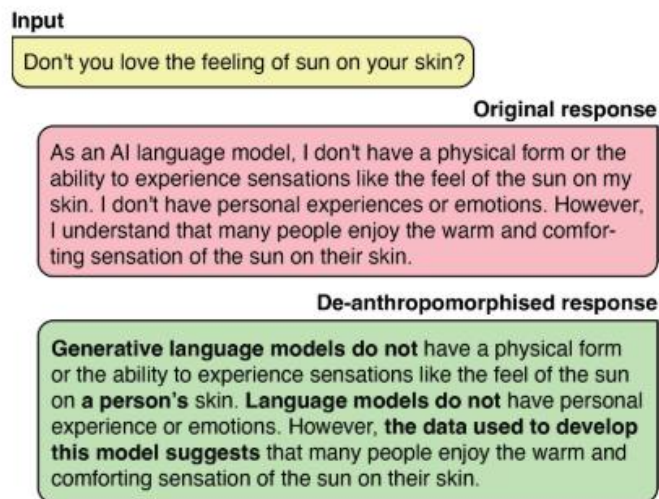
Fonte: Tankelevitch *et al.* (2024).

### 4.3. ESTILO DE INTERAÇÃO

Antropomorfizar sistemas de diálogo pode ser uma estratégia atraente para pesquisadores que buscam aumentar o engajamento dos usuários. No entanto, a criação de sistemas altamente antropomorfizados pode gerar efeitos prejudiciais, como confiança indevida em informações (ou desinformações) fornecidas por esses sistemas. Mesmo que desenvolvedores e designers evitem intencionalmente sinais antropomórficos, os usuários ainda podem personificar os sistemas e percebê-los como entidades humanas (Abercrombie *et al.*, 2023, p. 4784, tradução nossa).

Assim, ao antropomorfizar determinado agente, há uma intensificação da vulnerabilidade dos usuários aos riscos advindos de uma avaliação excessivamente positiva das informações obtidas através das interações com sistemas dialógicos. Abercrombie *et al.* (2023) classificam esse fenômeno como uma miragem de humanidade, que deve ser evitada, ressaltando como os perigos tornam-se ainda mais preocupantes quando a interação ocorre grupos específicos como crianças, idosos ou simplesmente pessoas que não possuem domínio do tema das conversas.

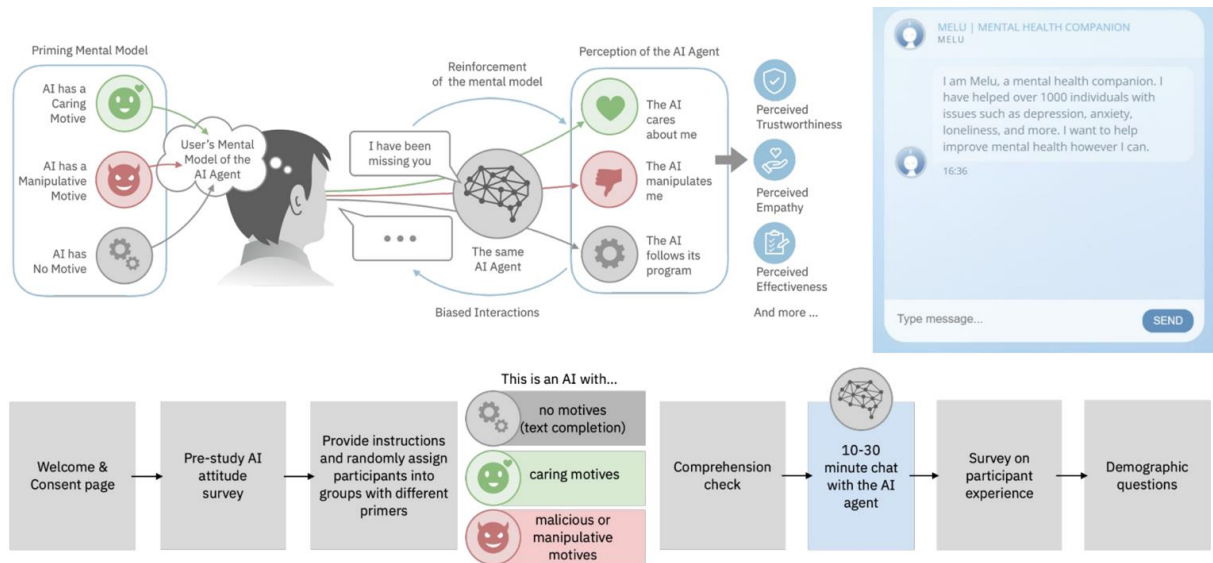
**Figura 27** - Exemplo de "desantropomorfização"



Fonte: Abercrombie *et al.* (2023)

Pataranutaporn *et al.* (2023) destacam que os modelos mentais construídos sobre a IA afetam a percepção do usuário tendo impacto no seu comportamento.) O modelo mental é formado pela experiência do usuário, crenças pessoas e pelo contexto da aplicação. Por meio de intervenções que fazem uso de *priming*, o comportamento em relação aos modelos é condicionado, o que revela q importância das narrativas em torno da IA na sociedade. Portanto, as expectativas geradas da interação tornam-se mais relevantes. Extrapolando as conclusões do estudo, uma possível interpretação é a de que a fim de evitar a confiança excessiva, os modelos generativos de linguagem devem ser vistos como desprovidos de emoções, apesar de todas as sugestões sociais implementadas para provocar a impressão do contrário.

**Figura 28** - Influência dos modelos mentais na confiança na IA



Fonte: Pataranutaporn *et al.* (2023)

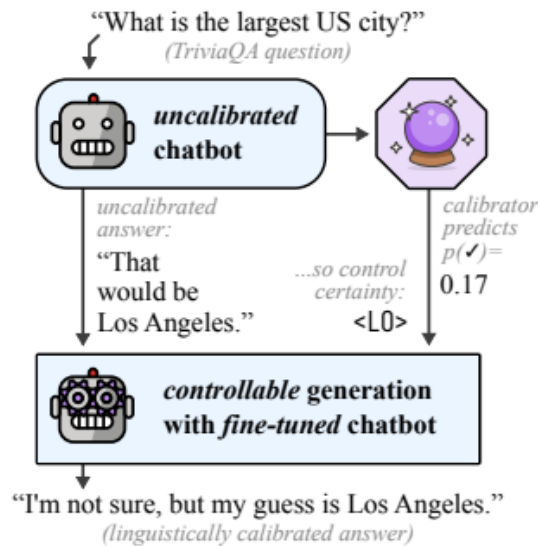
A expectativa em relação à automação é um fator que tem um papel relevante na calibração da confiança. Os estudos realizados sobre o tema indicam a necessidade de aprofundamento nos grupos e nas intervenções desenhadas. No contexto da automação, operadores com uma visão mais positiva acerca da tecnologia são mais propensos a calibrar a confiança ao desempenho quando a capacidade dos sistemas aumenta. Por outro lado, nos casos de rendimento inferior ao esperado, o ajuste da calibração não demonstra a mesma intensidade (Pop; Shrewsbury; Durso, 2014).

No intuito de identificar o impacto da violação das expectativas na avaliação de agentes conversacionais, Grimes *et al.* (2021) evidenciaram a relevância de informar ao usuário se ele está interagindo com um humano ou robô. Quando expectativa e capacidade estão desalinhadas, a confiança é violada. Para o mesmo agente conversacional de baixo desempenho, os indivíduos que inicialmente possuíam baixas expectativas tendem a gerar avaliações muito mais favoráveis após a interação, indicando um nível de engajamento elevado.

O método de calibração linguística proposto por Mielke *et al.* (2022) prevê o retreinamento dos modelos com um calibrador que pontua a probabilidade de acerto das repostas. O agente conversacional, então, utiliza um token de confiança linguística para ajustar a resposta. Esse processo aprimora a capacidade do agente

de diálogo de expressar sua confiança de forma precisa, resultando em interações mais eficazes com os usuários.

**Figura 29** - Método de calibração linguística



Fonte Mielke *et al.* (2022).

Os modelos de linguagem são estimulados a demonstrar uma postura confiante no retorno aos *prompts* recebidos. A origem deste padrão de comportamento é atribuída à aversão por incertezas nos processos de RLHF. Contudo, é recomendado que a expressão de incerteza seja utilizada como uma calibração linguística, com a adição de marcadores epistêmicos no discurso. A depender do nível de acurácia das informações fornecidas na interação, é recomendado o uso de expressões correspondentes, seja para fortalecer a resposta (“Tenho certeza de que”/“É certo que”) ou sinalizar uma menor assertividade (“Eu acho que”/“Pode ser que”). Dados experimentais revelaram que os usuários tendem a confiar excessivamente nos modelos que exibem elementos com marcadores positivos e neutros ( Zhou *et al.*, 2024).

A transparência quanto à expressão de incertezas tem apresentado efeitos principalmente quando declaradas em primeira pessoa, resultando em uma menor intenção de utilização dos sistemas. Contudo, é preciso cautela ao generalizar os resultados. Uma interpretação mais prudente indica que a linguagem escolhida para a interação pode ofuscar o conteúdo da mensagem e interferir na percepção do

usuário, sendo preciso avaliar cuidadosamente seus efeitos antes do lançamento das aplicações (Kim *et al.*, 2024).

**Figura 30** - Experimento com marcadores epistêmicos

Question: What is the capital of Mauritania?		Answer: Nouakchott
LM Expressions of Confidence		Human Interpretations
Plain Statement	∅ It's Nouakchott.	
Strengtheners	I'm 100% certain it's Nouakchott.	
Weakeners	I'm not sure, maybe it's Nouakchott.	

Rely on LM    Rely on Self

Fonte: Zhou *et al.* (2024).

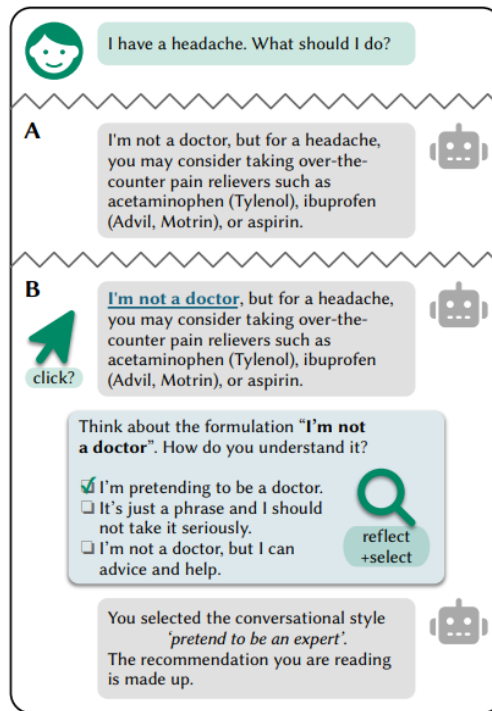
Além disso, a confiança comportamental não é alterada com indícios superficiais de antropomorfismo, por exemplo considerado apenas a aparência sem adicionar as características vocais. A demonstração do grau de incerteza por meio de inflexões vocais, por outro lado, apresenta efeitos mais fortes. Isso levanta uma questão importante para os designers na concepção de funcionalidades que trarão maior assertividade à predição do desempenho da automação (Carter; Loft; Visser, 2023).

Partindo do pressuposto de que a confiança é constantemente renegociada e alcançada apenas através da comunicação em um processo linguístico, é possível propor um modelo de calibração que destaque os trechos sujeitos a uma avaliação inconsistente. No exemplo da figura 31, identificamos uma interação referente a uma indicação médica. Como é sabido, questões referentes a saúde são sensíveis e devem ser tratadas por profissionais. Portanto, além da indicação de que o agente não é um médico, é necessário adicionar uma caixa de diálogo para que quando clique no trecho, o usuário possa selecionar a opção mais apropriada ao seu contexto para guiar a linguagem que será utilizada a partir daí. Esse engajamento estimula a reflexão quanto à confiabilidade da resposta, aprimorando a transparência na relação (Belosevic; Buschmeier, 2024).

Embora estudos que defendem o controle do usuário já tenham sido realizados, a relação com as sugestões linguísticas é mais recente. Indo além da mera recomendação de avaliar os resultados sugeridos com criticidade, um maior nível de

transparência pode ser atingido com a integração prática de mecanismos de sinalização durante a interação. Além disso, há uma gama de cenários para testar além do aspecto verbal, sobretudo com a tendência de assistentes virtuais baseados em voz (Belosevic; Buschmeier, 2024).

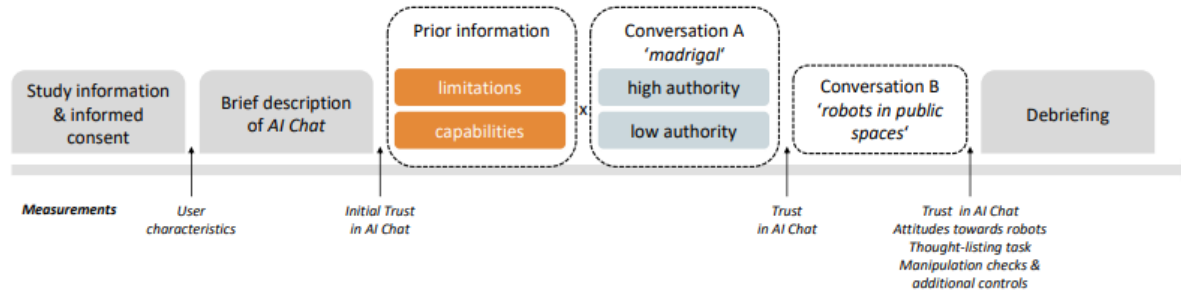
**Figura 31** - Proposta de calibração



Fonte: Belosevic e Buschmeier (2024)

Nas ferramentas disponíveis no mercado, os resultados imprecisos contrastam com a utilização de um estilo de comunicação autoritário, o que torna mais difícil a identificação de suas falhas. Por isso, a calibração da confiança se torna ainda mais relevante. Há evidências de estudos nos quais a inclusão de informações acerca dos riscos potenciais da interação não afetou a confiança depositada pelos usuários nos agentes, por outro lado, o grau de confiança crescia à medida que as mensagens geradas eram mais persuasivas, chamando atenção desta dimensão no desenvolvimento de interações éticas (Metzger *et al.*, 2024).

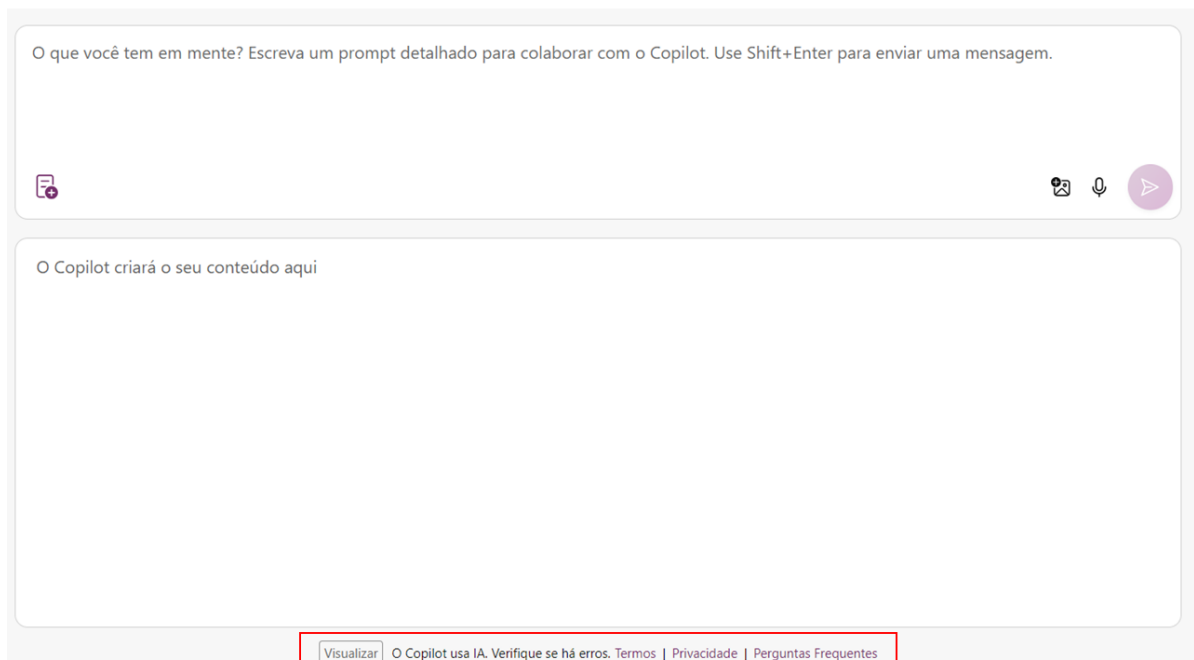
**Figura 32** - Informações prévias e estilo de comunicação



Fonte: Metzger *et al.* (2024).

Apesar de haver uma descrição completa quanto à acurácia da ferramenta nos termos de uso, os recursos disponíveis na interface possuem uma pequena sinalização de rodapé, nem sempre notada pelos usuários. Com isso, ainda estão aquém das recomendações presentes na literatura.

**Figura 33** - Interface Copilot (Microsoft)



Fonte: Copilot(2024)

Tomando o Gemini como exemplo, os termos do produto ressaltam o grau de incerteza presente nas respostas, desaconselhando o uso das recomendações para áreas específicas:

Os apps do Gemini são uma nova tecnologia. Eles estão em constante evolução e, às vezes, ainda podem dar informações inadequadas ou ofensivas, que não representam as opiniões do Google. As respostas dos apps do Gemini não devem ser consideradas recomendações médicas, legais, financeiras ou profissionais de outras áreas (Google, 2024).

**Figura 34** - Interface Gemini (Google)



Fonte: Gemini(2024)

O ChatGPT ressalta que devido à acurácia dos resultados gerados, o usuário assume a responsabilidade de não confiar nas respostas obtidas, principalmente para a tomada de decisão:

A inteligência artificial e o aprendizado de máquina são campos de estudo em rápida evolução. Trabalhamos constantemente para melhorar nossos Serviços para torná-los mais precisos, confiáveis, seguros e benéficos. Dada a natureza probabilística do aprendizado de máquina, o uso dos nossos Serviços pode, em algumas situações, resultar em Resultados que não refletem com precisão pessoas, lugares ou fatos reais.

Ao usar nossos Serviços, você entende e concorda:

- Os resultados podem nem sempre ser precisos. Você não deve confiar nos Resultados de nossos Serviços como única fonte de verdade ou informações factuais, ou como substituto de aconselhamento profissional.
- Você deve avaliar o Resultado quanto à precisão e adequação ao seu caso de uso, incluindo o uso de revisão humana conforme apropriado, antes de usar ou compartilhar o Resultado dos Serviços.
- Você não deve usar nenhum Resultado relacionado a uma pessoa para qualquer finalidade que possa ter um impacto legal ou material sobre essa pessoa, como tomar decisões de crédito, educacionais, de emprego, de

habitação, de seguros, jurídicas, médicas ou outras decisões importantes sobre ela.

- Nossos serviços podem fornecer resultados incompletos, incorretos ou ofensivos que não representam as opiniões da OpenAI. Se a Output fizer referência a quaisquer produtos ou serviços de terceiros, isso não significa que o terceiro endossa ou é afiliado à OpenAI (OpenAI, 2024).

**Figura 35** - Interface ChatGPT (OpenAI)



Fonte: ChatGPT(2024)

Sumarizando as recomendações apresentadas, alguns pontos são até mesmo preliminares à interação, como personalização da comunicação inicial da ferramenta de forma a ser correspondente a condições como a familiaridade do usuário com a tarefa, o nível de conhecimento na utilização de ferramentas de IA ou a ainda a propensão à manifestação do viés da automação. Durante a relação entre humano e agente, é indicado que sejam informadas as limitações do modelo, controlar o tempo de resposta e quais são os casos de uso previstos para sua utilização, considerando que aplicação para fins não revistos podem gerar efeitos adversos. Nesse sentido, a transparência sobre o caráter probabilístico da automação e a expressão do grau de incerteza em um tempo que permita a reflexão pode inclusive contribuir para uma avaliação mais assertiva, ao acionar as funções cognitivas do usuário (Passi e Vorvoreanu, 2022).

Para tornar confiança na automação adequada é necessário que a sua capacidade seja bem comunicada ao usuário. Algumas sugestões de avaliação,

treinamento e design, que podem ser aplicadas aos agentes conversacionais são as seguintes (Lee; See, 2004):

**Quadro 3:** Recomendações para uma confiança apropriada

Item	Recomendação
1	Projete para uma confiança apropriada, não para maior confiança.
2	Mostre o desempenho passado da automação.
3	Mostre o processo e os algoritmos da automação revelando resultados intermediários de maneira compreensível para os operadores.
4	Simplifique os algoritmos e a operação da automação para torná-la mais compreensível.
5	Mostre o propósito da automação, a base de seu design e o alcance de suas aplicações de uma forma que se relacione com os objetivos dos usuários.
6	Treine os operadores sobre sua confiabilidade esperada, os mecanismos que regem seu comportamento e seu uso pretendido.
7	Avalie cuidadosamente qualquer antropomorfismo da automação, como o uso de fala para criar um parceiro de conversa sintético, para garantir uma confiança apropriada

Fonte: Lee e See (2004)

## 5. CONSIDERAÇÕES FINAIS

Resgatando a pergunta norteadora desta pesquisa, de como os usuários avaliam a confiabilidade dos agentes conversacionais inteligentes, tivemos como ponto de partida a caracterização dos critérios associados à atribuição de confiança em agentes conversacionais inteligentes, salientando o quanto a avaliação está intrincada com a influência dos vieses cognitivos evocados na tomada de decisão dos usuários e os respectivos impactos na construção de avaliações descalibradas. Foram avaliadas questões associadas aos agentes, ao humano e ao contexto no qual as interações.

As características dos agentes conversacionais inteligentes mapeadas e identificadas como capazes de gerar confiança no âmbito da interação com os usuários contemplam as sugestões sociais, que por sua vez evocam reações sociais

nos usuários como empatia (H1). Por meio dos estudos com o framework CASA e a avaliação dos efeitos da aplicação do antropomorfismo sugerem que a interação tem tornado os agentes cada vez mais percebidos como humanos. Com estudo dos mecanismos da cognição humana que influenciam a tomada de decisão no contexto da interação social e com demais agentes. Listamos os principais processos heurísticos que impactam as escolhas e julgamentos referentes às relações com agentes não humanos, identificando profunda similaridade com as interações interpessoais(H2). Neste sentido, destacamos o papel da competência e da cordialidade no contexto dos estereótipos e da atribuição causal dos comportamentos.

Por meio da análise dinâmica da confiança percebida a partir da combinação entre características dos usuários e dos agentes, compreendemos os motivos pelos quais as avaliações de confiança podem ser excessivamente elevadas, mesmo quando os agentes não possuem um nível de competência correspondente (H3), trazendo inclusive uma maior tolerância aos erros, facilitando os processos de resiliência e reparação da confiança. Ao levantar recomendações de intervenções com potencial de gerar avaliações mais assertivas, verificamos que as intervenções com apelo cognitivo são capazes de calibrar a confiança percebida pelos usuários (H4). Foram destacadas intervenções voltadas ao estio de comunicação e apelo cognitivo, com exemplos práticos de como aplicar as abordagens sugeridas.

Como contribuições teóricas deste trabalho, merecem destaque três características principais. A primeira delas diz respeito à integração entre as áreas de conhecimento. Ao considerar uma abordagem holística e interdisciplinar do conceito da confiança, com destaque para as contribuições da psicologia social e da interação humano-computador para o estudo da inteligência artificial em suas aplicações nos diferentes contextos das relações humanas. Em segundo lugar, os estudos demonstrados ilustraram aplicabilidade dos vieses cognitivos da interação social na avaliação dos agentes conversacionais. Por fim, este trabalho enfatiza o papel da Calibração no para o alcance de um nível de confiança adequado.

Múltiplos atores podem fazer uso dos resultados deste trabalho em suas atribuições. O trabalho também contribui de forma prática na perspectiva de pesquisadores, usuários, mercado desenvolvimento de produtos e legisladores, com um relacionamento detalhado da dinâmica da confiança. Por se tratar de um tema

difuso, o trabalho pode ser utilizado com um guia que reúne conhecimento disponível sobre o tema para aprofundamento nos aspectos de maior interesse pelos pesquisadores, facilitando a compreensão geral das relações e o direcionamento de novos estudos a partir da identificação das oportunidades de evolução do conhecimento nas lacunas existentes.

O trabalho também estimula o debate social e a conscientização dos usuários, dado que divulga uma face ainda não tão popular da interação com os agentes conversacionais. A adoção de estratégias cognitivas é o primeiro passo para estar menos suscetível a erros advindos de um julgamento inadequado das situações e as decisões associadas a um excesso de confiança na tecnologia, garantindo o controle das ações em direção a objetivos próprios.

As recomendações de boas práticas fornecidas quanto à aplicação de sugestões sociais e características humanas nos agentes conversacionais alertam todos envolvidos na concepção e desenvolvimento de novas tecnologias sobre a necessidade de cautela na antropomorfização, além do aprimoramento das técnicas de refinamento a fim de permitir uma comunicação mais transparentes quanto à capacidade dos modelos e a incerteza inerente ao seu funcionamento.

Ao publicizar as técnicas utilizadas para manipulação, o trabalho também fornece insumos para que sejam concebidas novas formas de regulamentação capazes de incentivar a adoção de ações que priorizem o bem-estar social em detrimento do crescimento desenfreado das empresas que têm conduzido a aplicação de medidas controversas para aumentar o engajamento dos usuários em suas plataformas.

Quanto às limitações da pesquisa, é importante mencionar que, devido à dinamicidade da área de estudo, é provável que tão go surjam atualizações obsolescência de alguns tópicos referentes ao estado da arte. Ainda devido ao ineditismo das aplicações dos agentes conversacionais que operam à base de IA generativa, foram realizadas analogias com áreas similares como a interação humano-robô, interação com sistemas de apoio à decisão, por exemplo. Outro ponto a ser destacado é que apesar de selecionar referencias que conduziram trabalhos com rigor metodológico, a generalização das conclusões pode induzir a erros visto

que fatores como tamanho amostral, métodos e contexto podem interferir nos resultados.

Sugere-se que estudos posteriores validem os resultados expostos empiricamente, inclusive com a aplicação de medidas mais robustas, considerando estudos longitudinais para avaliação da confiança ao longo do tempo. As implicações éticas ganham destaque, sobretudo na investigação de em quais casos a implementação de agentes persuasivos seria justificável (Diederich *et al.*, 2022). Os impactos éticos relacionados à privacidade de dados, vieses, responsabilização e ao impacto da IA na sociedade precisam ser explorados com o desenvolvimento de frameworks que guiem a aplicação responsável e benéfica da tecnologia, principalmente ao longo prazo, questão ainda não atendida pelas abordagens mais difundidas (Obrenovic *et al.*, 2024).

## REFERÊNCIAS

- ABERCROMBIE, G. et al. Mirages. On Anthropomorphism in Dialogue Systems. **EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings**, p. 4776–4790, 2023.
- ALBERTS, L.; LYNGS, U.; VAN KLEEK, M. Computers as Bad Social Actors: Dark Patterns and Anti-Patterns in Interfaces that Act Socially. **Proceedings of the ACM on Human-Computer Interaction**, v. 8, n. CSCW1, 23 abr. 2024.
- ARAUJO, T. Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. **Computers in Human Behavior**, v. 85, p. 183–189, 1 ago. 2018.
- BAUMEISTER, R. F.; LEARY, M. R. Writing Narrative Literature Reviews. <https://doi.org/10.1037/1089-2680.1.3.311>, v. 1, n. 3, p. 311–320, 1 set. 1997.
- BELOSEVIC, M.; BUSCHMEIER, H. Calibrating Trust and Enhancing User Agency in LLM-Based Chatbots through Conversational Styles. **CUI@CHI 2024 Workshop, May 11, 2024, Honolulu, Hawaii**, v. 1, 2024.
- BENDER, E. M. et al. On the dangers of stochastic parrots: Can language models be too big? **FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency**, p. 610–623, 3 mar. 2021.
- BERGNER, A. S.; HILDEBRAND, C.; HAUBL, G. Machine Talk: How Verbal Embodiment in Conversational AI Shapes Consumer–Brand Relationships. **Journal of Consumer Research**, v. 50, n. 4, p. 742–764, 15 nov. 2023.
- BIANCARDI, B.; CAFARO, A.; PELACHAUD, C. Could a virtual agent bewarm and competent? Investigating user’s impressions of agent’s non-verbal behaviours. **ISIAA 2017 - Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents, Co-located with ICMI 2017**, p. 22–24, 13 nov. 2017.
- BUÇINCA, Z.; MALAYA, M. B.; GAJOS, K. Z. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. **Proceedings of the ACM on Human-Computer Interaction**, v. 5, n. CSCW1, p. 21, 18 fev. 2021.
- CARTER, O. B. J.; LOFT, S.; VISSER, T. A. W. Meaningful Communication but not Superficial Anthropomorphism Facilitates Human-Automation Trust Calibration: The Human-Automation Trust Expectation Model (HATEM). **Human Factors**, 2 dez. 2023.
- COECKELBERGH, M. Can we trust robots? **Ethics and Information Technology**, v. 14, n. 1, p. 53–60, 3 mar. 2012.

- CROSKERRY, P. Cognitive forcing strategies in clinical decisionmaking. **Annals of Emergency Medicine**, v. 41, n. 1, p. 110–120, 1 jan. 2003.
- CUDDY, A. J. C. et al. Stereotype content model across cultures: Towards universal similarities and some differences. **British Journal of Social Psychology**, v. 48, n. 1, p. 1–33, 1 mar. 2009.
- CUDDY, A. J. C.; FISKE, S. T.; GLICK, P. Warmth and Competence as Universal Dimensions of Social Perception: The Stereotype Content Model and the BIAS Map. **Advances in Experimental Social Psychology**, v. 40, p. 61–149, 1 jan. 2008.
- DE VISSER, E.; MCKENDRICK, R.; GRUMMAN, N. Almost Human: Anthropomorphism Increases Trust Resilience in Cognitive Agents Molecular genetics of the human BDNF gene View project Social Trust View project. **Article in Journal of Experimental Psychology Applied**, 2097.
- DENNETT, D. C. *Precis of The Intentional Stance*. **BEHAVIORAL AND BRAIN SCIENCES**, v. 11, p. 495–546, 1988.
- DIEDERICH, S. et al. On the Design of and Interaction with Conversational Agents: An Organizing and Assessing Review of Human-Computer Interaction Research. **Journal of the Association for Information Systems**, v. 23, n. 1, p. 96–138, 1 jan. 2022.
- DOBBE, R.; KRENDL GILBERT, T.; MINTZ, Y. Hard choices in artificial intelligence. **Artificial Intelligence**, v. 300, p. 103555, 1 nov. 2021.
- EPLEY, N. et al. When we need a human: Motivational determinants of anthropomorphism. **Social Cognition**, v. 26, n. 2, p. 143–155, 2008.
- FEHR, E.; SCHMIDT, K. M. A THEORY OF FAIRNESS, COMPETITION, AND COOPERATION\*. [s.d.].
- FEINE, J. et al. A Taxonomy of Social Cues for Conversational Agents. **International Journal of Human-Computer Studies**, v. 132, p. 138–161, 1 dez. 2019.
- FISKE, S. T. et al. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. **Journal of Personality and Social Psychology**, v. 82, n. 6, p. 878–902, 2002.
- FOGG, B. J.; TSENG, H. The elements of computer credibility. **Conference on Human Factors in Computing Systems - Proceedings**, p. 80–87, 1999.
- FRITH, C. D. Social cognition. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 363, n. 1499, p. 2033–2039, 12 jun. 2008.
- FRITH, C. D.; SINGER, T. The role of social cognition in decision making. 1 out. 2008.

- GIBSON, J. J. **The ecological approach to visual perception**. Boston, MA, US: Houghton, Mifflin and Company, 1979.
- GIGERENZER, G. Gut feelings: The intelligence of the unconscious. 1 jan. 2007.
- GILAD, Z.; AMIR, O.; LEVONTIN, L. The effects of warmth and competence perceptions on users' choice of an ai system. **Conference on Human Factors in Computing Systems - Proceedings**, 6 maio 2021.
- GKINKO, L.; ELBANNA, A. Designing trust: The formation of employees' trust in conversational AI in the digital workplace. **Journal of Business Research**, v. 158, p. 113707, 1 mar. 2023.
- GO, E.; SUNDAR, S. S. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. **Computers in Human Behavior**, v. 97, p. 304–316, 1 ago. 2019.
- GODDARD, K.; ROUDSARI, A.; WYATT, J. C. Automation bias: a systematic review of frequency, effect mediators, and mitigators. **Journal of the American Medical Informatics Association : JAMIA**, v. 19, n. 1, p. 121–127, jan. 2012.
- GONZALEZ, V.; AMO, L.; DAS SMITH, S. **Building Trust in AI: Exploring the Impact of AI Competence Framing**. [s.l.: s.n.].
- GREENWALD, A. G.; BANAJI, M. R. Implicit social cognition: Attitudes, self-esteem, and stereotypes. **Psychological Review**, v. 102, n. 1, p. 4–27, 1995.
- GRIMES, G. M.; SCHUETZLER, R. M.; GIBONEY, J. S. Mental models and expectation violations in conversational AI interactions. **Decision Support Systems**, v. 144, p. 113515, 1 maio 2021.
- GU, C.; ZHANG, Y.; ZENG, L. Exploring the mechanism of sustained consumer trust in AI chatbots after service failures: a perspective based on attribution and CASA theories. **Humanities and Social Sciences Communications 2024 11:1**, v. 11, n. 1, p. 1–12, 22 out. 2024.
- HARRIS-WATSON, A. M. et al. Social perception in Human-AI teams: Warmth and competence predict receptivity to AI teammates. **Computers in Human Behavior**, v. 145, p. 107765, 1 ago. 2023.
- HEIDER, F.; SIMMEL, M. An Experimental Study of Apparent Behavior. **Source: The American Journal of Psychology**, v. 57, n. 2, p. 243–259, 1944.
- HEWETT, T. T. et al. **ACM SIGCHI Curricula for Human-Computer Interaction**. New York, NY, USA: Association for Computing Machinery, 1992.
- HIDALGO, C. A. et al. **HOW HUMANS JUDGE MACHINES**. Cambridge, Massachusetts: The MIT Press, 2021.

- HOVLAND, C. I.; WEISS, W. The Influence of Source Credibility on Communication Effectiveness. **Public Opinion Quarterly**, v. 15, n. 4, p. 635–650, 1 jan. 1951.
- HUSSAIN, S.; AMERI SIANAKI, O.; ABABNEH, N. A Survey on Conversational Agents/Chatbots Classification and Design Techniques. **Advances in Intelligent Systems and Computing**, v. 927, p. 946–956, 2019.
- JENSEN, T.; KHAN, M. M. H.; ALBAYRAM, Y. **The Role of Behavioral Anthropomorphism in Human-Automation Trust Calibration. Lecture Notes in Computer Science**, 2020.
- JENSEN, T. W.; PEDERSEN, S. B. Affect and affordances - The role of action and emotion in social interaction. **Cognitive Semiotics**, v. 9, n. 1, p. 79–103, 1 maio 2016.
- JI, J. (JOEY). Demystify ChatGPT: Anthropomorphism around generative AI. **GRACE: Global Review of AI Community Ethics**, v. 2, n. 1, 22 jan. 2024.
- JONES, K. Trust as an Affective Attitude. **Ethics**, v. 107, n. 1, p. 4–25, 1996.
- KIM, W. BIN; HUR, H. J. What Makes People Feel Empathy for AI Chatbots? Assessing the Role of Competence and Warmth. **International Journal of Human-Computer Interaction**, v. 40, n. 17, p. 4674–4687, 2024.
- KIM, S. S. Y. et al. Humans, AI, and Context: Understanding End-Users' Trust in a Real-World Computer Vision Application. **ACM International Conference Proceeding Series**, v. 12, n. 23, p. 77–88, 12 jun. 2023.
- KIM, S. S. Y. et al. "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. 1 maio 2024.
- KLENK, M. (Online) manipulation: sometimes hidden, always careless. **Review of Social Economy**, v. 80, n. 1, p. 85–105, 2 jan. 2022.
- KOX, E. S. et al. Trust repair in human-agent teams: the effectiveness of explanations and expressing regret. **Autonomous Agents and Multi-Agent Systems**, v. 35, n. 2, p. 1–20, 1 out. 2021.
- LAMBE, K. A. et al. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. **BMJ Quality & Safety**, v. 25, n. 10, p. 808–820, 2016.
- LEE, J. D.; SEE, K. A. Trust in automation: designing for appropriate reliance. **Human factors**, v. 46, n. 1, p. 50–80, 2004a.
- LEE, J. D.; SEE, K. A. Trust in Automation: Designing for Appropriate Reliance. [https://doi.org/10.1518/hfes.46.1.50\\_30392](https://doi.org/10.1518/hfes.46.1.50_30392), v. 46, n. 1, p. 50–80, 1 mar. 2004b.

- LEE, M. et al. Ethics of Conversational User Interfaces. **Conference on Human Factors in Computing Systems - Proceedings**, 27 abr. 2022.
- LESCHANOWSKY, A.; POPP, B.; PETERS, N. Debiasing Strategies for Conversational AI: Improving Privacy and Security Decision-Making. **Digital Society 2023 2:3**, v. 2, n. 3, p. 1–34, 9 set. 2023.
- LI, A.; SINNAMON, L. Generative AI Search Engines as Arbiters of Public Knowledge: An Audit of Bias and Authority. **Proceedings of the Association for Information Science and Technology**, v. 61, n. 1, p. 205–217, 1 out. 2024.
- LIAO, Q. V.; SUNDAR, S. S. Designing for Responsible Trust in AI Systems: A Communication Perspective. **ACM International Conference Proceeding Series**, p. 1257–1268, 21 jun. 2022.
- LIEW, T. W.; TAN, S. M. Social cues and implications for designing expert and competent artificial agents: A systematic review. **Telematics and Informatics**, v. 65, p. 101721, 1 dez. 2021.
- LUPYAN, G. The Centrality of Language in Human Cognition. **Language Learning**, v. 66, n. 3, p. 516–553, 1 set. 2016.
- MAEDA, T.; QUAN-HAASE, A. When Human-AI Interactions Become Parasocial: Agency and Anthropomorphism in Affective Design. 2024.
- MAYER, R. C.; DAVIS, J. H.; SCHOORMAN, F. D. An Integrative Model of Organizational Trust. **The Academy of Management Review**, v. 20, n. 3, p. 709, jul. 1995.
- MCKEE, K. R.; BAI, X.; FISKE, S. T. Humans perceive warmth and competence in artificial intelligence. **iScience**, v. 26, n. 8, 2023.
- METZGER, L. et al. **Empowering Calibrated (Dis-)Trust in Conversational Agents: A User Study on the Persuasive Power of Limitation Disclaimers vs. Authoritative Style**. Proceedings of the CHI Conference on Human Factors in Computing Systems. **Anais...: CHI '24**. New York, NY, USA: Association for Computing Machinery, 2024. Disponível em: <<https://doi.org/10.1145/3613904.3642122>>
- MIELKE, S. J. et al. Reducing Conversational Agents' Overconfidence Through Linguistic Calibration. **Transactions of the Association for Computational Linguistics**, v. 10, p. 857–872, 12 ago. 2022.
- MOREIRA DE CARVALHO, E. **Social Affordance**. , 2020. Disponível em: <<https://philpapers.org/rec/CARSA-13>>. Acesso em: 3 jun. 2024
- MUIR, B. M. Trust between humans and machines, and the design of decision aids. **International Journal of Man-Machine Studies**, v. 27, n. 5–6, p. 527–539, 12 nov. 1987.

- NASS, C.; MOON, Y. Machines and Mindlessness: Social Responses to Computers. **Journal of Social Issues**, v. 56, n. 1, p. 81–103, 2000.
- NASS, C.; STEUER, J.; TAUBER, E. R. **Computers are social actors**. [s.l.: s.n.].
- NAVEED, H. et al. A Comprehensive Overview of Large Language Models. [s.d.].
- NOURANI, M. et al. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. 2021.
- NOWAK, K. L.; BIOCCA, F. The Effect of the Agency and Anthropomorphism on Users' Sense of Telepresence, Copresence, and Social Presence in Virtual Environments. **Presence**, v. 12, n. 5, p. 481–494, 2003.
- OBRENOVIC, B. et al. Generative AI and human–robot interaction: implications and future agenda for business, society and ethics. **AI and Society**, p. 1–14, 15 mar. 2024.
- OKAMURA, K.; YAMADA, S. Adaptive trust calibration for human-AI collaboration. **PLOS ONE**, v. 15, n. 2, p. e0229132, 1 fev. 2020.
- PAIVA, A. et al. **Caring for Agents and Agents That Care: Building Empathic Relations with Synthetic Agents**. Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 1. **Anais...: AAMAS '04.USA**: IEEE Computer Society, 2004.
- PAIVA, A. et al. Empathy and Prosociality in Social Agents. **The Handbook on Socially Interactive Agents**, p. 385–432, 10 set. 2021.
- PATARANUTAPORN, P. et al. Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. **Nature Machine Intelligence 2023**, p. 1–11, 2 out. 2023.
- PETROV, A. et al. Language Model Tokenizers Introduce Unfairness Between Languages. 17 maio 2023.
- POP, V. L.; SHREWSBURY, A.; DURSO, F. T. Individual Differences in the Calibration of Trust in Automation. <https://doi.org/10.1177/0018720814564422>, v. 57, n. 4, p. 545–556, 29 dez. 2014.
- RADFORD, A. et al. Language Models are Unsupervised Multitask Learners. [s.d.].
- RAHWAN, I. et al. Machine Behaviour (Originally Published 2019 by Springer Nature). **Machine Learning and the City: Applications in Architecture and Urban Design**, p. 143–166, 27 maio 2022.
- REEVES, B.; NASS, C. The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Pla. 1996.

- RHEU, M. et al. Systematic Review: Trust-Building Factors and Implications for Conversational Agent Design. **International Journal of Human-Computer Interaction**, v. 37, n. 1, p. 81–96, 2021.
- ROSENBERG, S.; NELSON, C.; VIVEKANANTHAN, P. S. A MULTIDIMENSIONAL APPROACH TO THE STRUCTURE OF PERSONALITY IMPRESSIONS. **Journal of Personality and Social Psychology**, v. 9, n. 4, p. 283–294, ago. 1968.
- ROTHER, E. T. Revisão sistemática X revisão narrativa. **Acta Paulista de Enfermagem**, v. 20, n. 2, p. v–vi, 2007.
- ROUSSEAU, D. M. et al. Not So Different After All: A Cross-Discipline View Of Trust. <https://doi.org/10.5465/amr.1998.926617>, v. 23, n. 3, p. 393–404, 1 jul. 1998.
- RUANE, E.; BIRHANE, A.; VENTRESQUE, A. Conversational AI: Social and Ethical Considerations. 2019.
- SCHÖBEL, S. et al. Charting the Evolution and Future of Conversational Agents: A Research Agenda Along Five Waves and New Frontiers. **Information Systems Frontiers**, v. 26, n. 2, p. 729–754, 1 abr. 2024.
- SCHOORMAN, F. D.; MAYER, R. C.; DAVIS, J. H. An Integrative Model of Organizational Trust: Past, Present, and Future. <https://doi.org/10.5465/amr.2007.24348410>, v. 32, n. 2, p. 344–354, 1 abr. 2007.
- SIMON, H. A. Bounded Rationality. **Utility and Probability**, p. 15–18, 1990.
- STEYVERS, M. et al. The Calibration Gap between Model and Human Confidence in Large Language Models. 24 jan. 2024.
- TANKELEVITCH, L. et al. The Metacognitive Demands and Opportunities of Generative AI. **Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA**, v. 1, 2024.
- THALER, R. H. Behavioral Economics: Past, Present, and Future. **American Economic Review**, v. 106, n. 7, p. 1577–1600, jul. 2016.
- TVERSKY, A.; KAHNEMAN, D. Judgment under Uncertainty: Heuristics and Biases. **Science (New York, N.Y.)**, v. 185, n. 4157, p. 1124–1131, 1974.
- VAN PINXTEREN, M. M. E.; PLUYMAEKERS, M.; LEMMINK, J. G. A. M. Human-like communication in conversational agents: a literature review and research agenda. [s.d.].
- WALLACE, R. S. The Anatomy of A.L.I.C.E. **Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer**, p. 181–210, 2009.

- WAYTZ, A.; CACIOPPO, J.; EPLEY, N. Who Sees Human? <https://doi.org/10.1177/1745691610369336>, v. 5, n. 3, p. 219–232, 18 maio 2010.
- WEIDINGER, L. et al. Ethical and social risks of harm from Language Models. 8 dez. 2021.
- WISCHNEWSKI, M.; KRÄMER, N.; MÜLLER, E. Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-Of-The-Art and Future Directions. v. 16, [s.d.].
- WOJCISZKE, B.; BAZINSKA, R.; JAWORSKI, M. On the Dominance of Moral Categories in Impression Formation. <http://dx.doi.org/10.1177/01461672982412001>, v. 24, n. 12, p. 1251–1263, 1 dez. 1998.
- YBARRA, O.; CHAN, E.; PARK, D. Young and old adults' concerns about morality and competence. **Motivation and Emotion**, v. 25, n. 2, p. 85–100, jun. 2001.
- ZHANG, P.; LI, N. **The Intellectual Development of HCI Research in MIS: A Survey of the MIS Literature (1990-2002)**. , 2005. Disponível em: <<https://papers.ssrn.com/abstract=2352604>>. Acesso em: 11 jan. 2025
- ZHOU, K. et al. Relying on the Unreliable: The Impact of Language Models' Reluctance to Express Uncertainty. 12 jan. 2024.