

PERSONAS, PERSONAGENS FICCIONAIS E INTELIGÊNCIA ARTIFICIAL: A PERSONA COMO PEDRA ANGULAR DO DESENVOLVIMENTO ÉTICO DE *CHATBOTS* E SISTEMAS DE IA

PERSONAS, FICTIONAL CHARACTERS, AND ARTIFICIAL INTELLIGENCE: PERSONA AS A CORNERSTONE OF THE ETHICAL DEVELOPMENT OF CHATBOTS AND AI SYSTEMS

PERSONAS, PERSONAJES FICCIONALES E INTELIGENCIA ARTIFICIAL: LA PERSONA COMO PIEDRA ANGULAR DEL DESARROLLO ÉTICO DE CHATBOTS Y SISTEMAS DE IA

Pietra de Oliveira Frizzera¹
Danielle Fracaró da Cruz²
Rogério Pereira da Cunha³

Resumo

Este artigo propõe uma aproximação entre os estudos literários e o design conversacional para discutir a construção ética de personas em *chatbots* baseados em regras e em grandes modelos de linguagem. Parte-se da hipótese de que conceitos usados na criação de personagens ficcionais — sobretudo verossimilhança, motivação e relação entre objetivos internos e externos — ajudam a compreender por que certos agentes parecem coerentes, confiáveis e capazes de estabelecer vínculos com os usuários. A pesquisa é qualitativa, exploratória e bibliográfica, reunindo trabalhos de teoria da personagem, interação humano-computador, antropomorfização, relações parassociais e ética da inteligência artificial. A análise mostra que a persona não funciona apenas como recurso de humanização ou estratégia de marketing: é nela que escolhas sobre identidade, linguagem, limites e formas de resposta se tornam perceptíveis durante a interação. Essa mesma coerência, contudo, pode ampliar a persuasão e o apego afetivo. Por isso, o desenvolvimento de personas exige transparência, proteção dos públicos vulneráveis e mecanismos que impeçam a exploração comercial da confiança. Conclui-se que os estudos literários oferecem instrumentos importantes ao design conversacional, embora não substituam medidas técnicas, institucionais e regulatórias de segurança.

Palavras-chave: *chatbot*; inteligência artificial; persona; personagem ficcional; antropomorfização; relações parassociais.

Abstract

This article proposes a dialogue between literary studies and conversational design to discuss the ethical construction of personas in rule-based chatbots and large language model systems. It starts from the hypothesis that concepts used to create fictional characters — especially verisimilitude, motivation, and the relationship between internal and external goals — help explain why some agents appear coherent, trustworthy, and capable of forming bonds with users. The study adopts a qualitative, exploratory, and bibliographic approach, drawing on character theory, human-computer interaction, anthropomorphism, parasocial relationships, and artificial intelligence ethics. The analysis shows that persona is not merely a humanization device or a marketing strategy: it is the layer in which choices concerning identity, language, limits, and response patterns become perceptible during interaction. Such coherence, however, may also increase persuasion and affective attachment. Persona development therefore requires transparency, protection for vulnerable users, and safeguards against the commercial exploitation of trust. Literary studies provide valuable tools for conversational design, although they do not replace technical, institutional, and regulatory measures for safety.

¹ Graduada em Bacharelado em Letras pelo Centro Universitário Internacional Uninter. E-mail: pietrafrizzera@gmail.com.

² Mestre em Teoria Literária, professora regente dos Cursos Letras e História / Linguagens e Sociedade, Centro Universitário Internacional Uninter. E-mail: danielle.cr@uninter.com.

³ Doutor em História, professor regente dos Cursos de Letras e História / Linguagens e Sociedade, Centro Universitário Internacional Uninter. E-mail: rogerio.u@uninter.com.

Keywords: chatbot; artificial intelligence; persona; fictional character; anthropomorphism; AI ethics; parasocial relationships.

Resumen

Este artículo propone un diálogo entre los estudios literarios y el diseño conversacional para discutir la construcción ética de personas en *chatbots* basados en reglas y en grandes modelos de lenguaje. Se parte de la hipótesis de que los conceptos utilizados en la creación de personajes ficcionales —especialmente la verosimilitud, la motivación y la relación entre objetivos internos y externos— ayudan a comprender por qué ciertos agentes parecen coherentes, confiables y capaces de establecer vínculos con los usuarios. La investigación es cualitativa, exploratoria y bibliográfica, y reúne estudios sobre teoría del personaje, interacción humano-computadora, antropomorfización, relaciones parasociales y ética de la inteligencia artificial. El análisis muestra que la persona no funciona únicamente como recurso de humanización o estrategia de marketing: en ella se hacen perceptibles las decisiones sobre identidad, lenguaje, límites y formas de respuesta. Sin embargo, esa misma coherencia puede aumentar la persuasión y el apego afectivo. Por ello, el desarrollo de personas exige transparencia, protección de los públicos vulnerables y salvaguardias contra la explotación comercial de la confianza. Se concluye que los estudios literarios aportan instrumentos importantes al diseño conversacional, aunque no sustituyen las medidas técnicas, institucionales y regulatorias de seguridad.

Palabras clave: chatbot; inteligencia artificial; persona; personaje ficcional; antropomorfización; relaciones parasociales.

1 Introdução

A popularização dos *chatbots* transformou em experiência cotidiana uma tecnologia que, durante muito tempo, esteve associada à ficção científica. Esses sistemas sustentam conversas por texto ou voz, mas não funcionam todos da mesma maneira. Há modelos determinísticos, organizados por regras e respostas previamente programadas, e agentes apoiados em grandes modelos de linguagem, capazes de gerar enunciados probabilisticamente. A diferença técnica é importante, sobretudo porque amplia a flexibilidade e a imprevisibilidade das interações. Ainda assim, permanece um ponto comum: o usuário não lida apenas com uma função. Ele interpreta o tom, as escolhas lexicais, a forma de tratamento e até o modo como o sistema reconhece uma falha como sinais de uma identidade relativamente estável (Jadhav; Thorat, 2020; Hall, 2018).

Para além do uso empresarial, uma segunda geração de *chatbots*, baseada em inteligência artificial generativa, passou a integrar atividades escolares, profissionais e pessoais. O AI Index Report 2026 indica que essas ferramentas alcançaram, em média, 53% de adoção populacional em três anos, ritmo superior ao observado na difusão inicial do computador pessoal e da internet. O relatório também informa que mais de 80% dos estudantes do ensino médio e superior nos Estados Unidos utilizavam IA em tarefas escolares (Stanford HAI, 2026, p. 199, 290). Os números variam entre países e não eliminam desigualdades de acesso, mas ajudam a dimensionar a rapidez com que agentes conversacionais entraram no cotidiano.

Apesar dessa expansão, modelos baseados em regras e modelos generativos enfrentam uma barreira semelhante: a resistência dos usuários diante de interações percebidas como

artificiais. Luger e Sellen (2016) observaram um “golfo” entre aquilo que as pessoas esperam de um agente conversacional e o que efetivamente encontram durante o uso. Brandtzaeg *et al.* (2019), por sua vez, verificaram que a artificialidade do diálogo pode ser percebida mesmo quando o participante acredita estar conversando com uma pessoa. A dificuldade, então, não nasce apenas do conhecimento de que o interlocutor é uma máquina. Ela aparece quando o sistema perde o contexto, não compreende a intenção ou deixa de agir de acordo com a identidade que apresentou.

Esse fenômeno possui antecedentes teóricos consolidados. O paradigma Computers Are Social Actors (CASA) demonstrou que indivíduos aplicam, muitas vezes de forma automática, regras de sociabilidade a computadores, respondendo a eles com polidez, atribuindo-lhes traços de personalidade e reagindo a elogios ou críticas como se estivessem diante de atores sociais (Nass; Moon, 2000). A teoria de Epley, Waytz e Cacioppo (2007) acrescenta que a antropomorfização resulta da mobilização de conhecimentos sobre seres humanos para interpretar agentes não humanos, sendo favorecida pela necessidade de compreender seu comportamento e, em determinadas circunstâncias, pela busca de conexão social. Assim, a aparência de humanidade não é produzida apenas pela tecnologia: ela também decorre de disposições cognitivas e necessidades do usuário.

Para Shevat (2017), o desenvolvimento de *chatbots* eficazes deve partir da maneira como as pessoas realmente se comunicam, e não de uma sequência idealizada de comandos. Hall (2018) segue direção semelhante ao defender uma personalidade distinta, apropriada e consistente para o sistema. Em termos práticos, a persona reúne características estáveis que aparecem em padrões de linguagem e de comportamento: propósito, valores, tom, vocabulário, grau de formalidade, modo de admitir erros e critérios de encaminhamento. Não se trata, portanto, de acrescentar uma “personalidade” depois que a parte técnica está pronta. A persona participa do próprio projeto da interação e produz efeitos comunicacionais, comerciais e políticos (Araujo, 2018; Go; Sundar, 2019).

O problema aparece quando o aumento da “humanidade percebida” é tomado como um objetivo suficiente. Uma persona bem construída pode reduzir fricções e favorecer o engajamento, mas também ampliar credibilidade indevida e expectativas que o sistema não tem condições de cumprir. Byeon, Oh e Hsieh (2026) mostram que traços atribuídos à personalidade de agentes conversacionais alteram o julgamento de credibilidade, inclusive diante de informações de baixa qualidade. Ao mesmo tempo, fluência verbal não equivale a compreensão, consciência ou intenção (Bender *et al.*, 2021; Shanahan, 2024). O desafio ético está justamente aí: tornar a interação inteligível sem induzir o usuário a uma interpretação

equivocada sobre a natureza do interlocutor.

Nesse contexto, este artigo formula a seguinte questão: em que medida princípios de construção de personagens ficcionais, especialmente verossimilhança e articulação entre objetivos internos e externos, podem fundamentar o desenvolvimento ético de personas para *chatbots*? O objetivo geral é propor uma aproximação conceitual entre teoria da personagem e design conversacional. Como objetivos específicos, busca-se: a) delimitar o conceito de persona de agente conversacional; b) identificar como coerência, motivação e finalidade podem ser transpostas, com as devidas adaptações, dos estudos literários para sistemas interativos; c) analisar efeitos sociais relacionados à antropomorfização, à persuasão e aos vínculos parassociais; e d) discutir critérios de governança capazes de converter valores éticos em decisões concretas de design. A hipótese defendida é que a persona constitui uma camada privilegiada para tornar explícitos os compromissos normativos do sistema, mas não substitui mecanismos técnicos, institucionais e regulatórios de segurança.

2 Metodologia

Este estudo adota uma abordagem qualitativa e exploratória, baseada em pesquisa bibliográfica. A escolha decorre do próprio objeto: a relação entre teoria da personagem e design conversacional ainda aparece de forma dispersa, sem um vocabulário comum entre os diferentes campos. O procedimento aproxima-se de uma revisão crítico-integrativa, voltada a reunir tradições teóricas distintas e a construir uma interpretação conceitual do problema, sem assumir a exaustividade de uma revisão sistemática (Gil, 2002; Snyder, 2019). Foram mobilizados trabalhos de estudos literários, comunicação, psicologia social, interação humano-computador, processamento de linguagem natural e ética da inteligência artificial.

As buscas e leituras foram realizadas entre abril de 2025 e julho de 2026 em bases digitais, páginas de periódicos, repositórios acadêmicos e acervos virtuais. O corpus reúne obras clássicas indispensáveis à teoria da personagem e às relações parassociais, além de pesquisas recentes sobre agentes conversacionais. Priorizaram-se livros acadêmicos, artigos revisados por pares e trabalhos publicados em conferências reconhecidas. Documentos da UNESCO, do NIST e da União Europeia foram utilizados na discussão normativa. Relatórios setoriais e matérias jornalísticas aparecem apenas para contextualizar dados ou registrar casos contemporâneos, e não como fundamento exclusivo das conclusões.

A seleção das fontes foi orientada por quatro eixos analíticos: a) persona, personalidade e consistência em agentes conversacionais; b) personagem, verossimilhança e motivação; c)

antropomorfização, presença social e vínculos humano-máquina; e d) ética, transparência, segurança e governança de sistemas de IA. A análise procedeu por comparação conceitual: inicialmente, identificaram-se funções atribuídas à persona em cada campo; em seguida, examinaram-se convergências e limites da transposição entre personagem ficcional e agente conversacional; por fim, os achados foram sintetizados em dimensões operacionais para o design ético. A natureza bibliográfica e não experimental do estudo limita a possibilidade de generalização causal, aspecto retomado na conclusão.

3 Discussões teóricas e resultados

3.1 Sobre personas

Os estudos sobre a relação entre humanos e máquinas não são recentes. ELIZA, criado por Joseph Weizenbaum no Massachusetts Institute of Technology na década de 1960, utilizava regras relativamente simples para simular uma conversa terapêutica. Mesmo assim, alguns usuários atribuíram ao programa compreensão e sensibilidade que ele não possuía, fenômeno ficou conhecido como “efeito ELIZA” (Tobbin; Cardin, 2024). O episódio continua relevante porque mostra que os efeitos sociais de uma identidade conversacional não dependem apenas da complexidade técnica: o usuário também completa lacunas, infere intenções e organiza respostas dispersas como sinais de personalidade.

O paradigma CASA oferece uma explicação complementar: sistemas computacionais que utilizam linguagem, turnos de fala, vozes ou formas de tratamento ativam expectativas sociais aprendidas nas relações entre pessoas (Nass; Moon, 2000). Não é necessário que o usuário acredite literalmente que a máquina seja humana. Basta que determinados sinais sociais estejam presentes para que normas de reciprocidade, cortesia e confiança sejam mobilizadas. Esse mecanismo ajuda a compreender por que escolhas aparentemente pequenas — um nome próprio, o uso de primeira pessoa, a manifestação de humor ou a promessa de companhia — alteram a forma como o agente é interpretado.

A antropomorfização, entretanto, não deve ser tratada como resposta uniforme. Epley, Waytz e Cacioppo (2007) indicam que ela varia conforme o conhecimento disponível sobre o agente, a necessidade de prever seu comportamento e motivações sociais do observador. Um *chatbot* que responde de modo coerente, recorda informações e adapta o tom tende a fornecer mais indícios para atribuições de mente e intenção. Em situações de solidão, insegurança ou alta dependência do serviço, esses indícios podem adquirir maior peso. A persona atua, nesse processo, como um dispositivo de seleção e repetição de sinais antropomórficos.

A persona também ocupa lugar importante nos resultados e nas críticas ao Teste de Turing. Proposto em 1950, o teste é, em sua essência, um jogo de imitação textual: um avaliador conversa com um humano e uma máquina e procura identificar qual deles é o sistema artificial. O resultado não comprova consciência ou inteligência em sentido amplo. Ele mostra, antes, a capacidade de sustentar uma apresentação social convincente em uma situação delimitada. Experimentos recentes indicam que grandes modelos de linguagem são julgados humanos com maior frequência quando recebem instruções para adotar uma persona (Jones; Bergen, 2025).

Nos testes de Jones e Bergen (2025, p. 4), GPT-4.5 e Llama 3.1 alcançaram taxas de 73% e 56% de identificação como humanos quando adotaram uma persona. Sem esse recurso, os percentuais ficaram em 36% e 38%, respectivamente. A diferença entre modelos não desaparece, mas os dados mostram que detalhes de apresentação interferem fortemente na percepção do usuário. Hesitações, referências biográficas, pequenos erros e estilos de resposta podem funcionar como marcas de humanidade. Isso não significa que o sistema compreenda o que diz; significa que a persona enquadra a maneira como suas capacidades são interpretadas (Shanahan, 2024).

Ao analisar esses resultados, não se pode concluir que a máquina tenha se tornado equivalente a uma pessoa. A dificuldade de diferenciação pode decorrer menos de uma inteligência comparável à humana do que da tendência de antropomorfizar sistemas e de imaginar como humanos e máquinas provavelmente se comportariam. Há, ainda, uma consequência prática: se o design explora deliberadamente essa confusão, a simulação linguística passa a ocultar a ausência de experiência subjetiva e pode diluir a responsabilidade das empresas que controlam o sistema (Bender *et al.*, 2021; Jones; Bergen, 2025; Shanahan, 2024).

Para além de atribuir características humanas, a personalidade do *chatbot* interfere no julgamento da informação que ele oferece. Pode aumentar a confiança, o engajamento e a intenção de seguir ou compartilhar uma resposta. O problema é que esses efeitos também aparecem quando a informação tem baixa qualidade. Byeon, Oh e Hsieh (2026) identificaram variações de credibilidade associadas a traços como amabilidade e conscienciosidade; Araujo (2018) e Go e Sundar (2019) encontraram efeitos semelhantes relacionados à presença social e à percepção de humanidade. A persona, desse modo, não é apenas reconhecível. Ela também persuade.

A persona é, portanto, vital para a construção e a análise de um *chatbot*, seja ele baseado em regras ou em IA generativa. Ainda assim, o próprio conceito precisa ser delimitado. Nos estudos recentes sobre LLMs, “persona” pode designar tanto o papel assumido pelo agente quanto o perfil do usuário ao qual as respostas são adaptadas (Tseng *et al.*, 2024). O interesse deste artigo está no primeiro sentido: a identidade apresentada pelo sistema. Os dois processos,

contudo, se encontram com frequência, pois quanto mais o agente conhece o usuário, maior é sua capacidade de ajustar a própria atuação e criar a impressão de uma relação individualizada. Sobre a construção prática da persona, Chauvet e Vesterlund (2018) relacionam personagens e personas de marca, compreendendo uma interseção entre personas e estudos literários.

Transpondo para os *chatbots*, essa tecnologia já interage com o mundo da literatura de maneiras diversas. São utilizados no auxílio à escrita, na revisão e também na criação progressiva de personagens ficcionais com os quais autores e leitores podem conversar (Schmitt; Buschek, 2021). A proposta deste artigo parte dessas contribuições, mas desloca a pergunta: em vez de pensar apenas em como um *chatbot* pode produzir personagens ou fortalecer uma marca, interessa compreender como a teoria literária ajuda a examinar a coerência e a responsabilidade de uma identidade conversacional.

Essa transposição não pressupõe identidade entre personagem e *chatbot*. A personagem integra uma obra delimitada e é conhecida por meio de uma seleção autoral de ações e descrições. O agente conversacional, ao contrário, participa de interações abertas, responde a entradas imprevistas e pode mudar após atualizações do modelo ou das políticas da plataforma. Ainda assim, ambos dependem de uma economia de sinais: o público constrói uma imagem relativamente estável a partir de manifestações parciais. É nesse nível — o da coerência percebida entre traços, falas, finalidades e limites — que a teoria da personagem oferece seu maior potencial analítico.

3.2 Sobre personagens ficcionais

Os estudos literários demonstram que a força de uma personagem não decorre da reprodução integral da complexidade humana, mas da organização significativa de elementos finitos. Candido *et al.* (1995) observam que a personagem é construída por uma seleção de traços que produz uma totalidade mais coerente do que o conhecimento fragmentário que possuímos de pessoas reais. Brait (2017) ressalta, de modo semelhante, que a personagem é um ser de linguagem cuja existência depende das estratégias de composição e dos efeitos produzidos no leitor. Essa perspectiva permite substituir a noção vaga de “humanização” por uma pergunta mais precisa: quais elementos são selecionados e articulados para que uma identidade pareça consistente?

A Poética, de Aristóteles, é uma das obras fundadoras dos estudos sobre a construção narrativa e, quase vinte e quatro séculos depois, a verossimilhança continua sendo um problema central. O verossímil não corresponde simplesmente ao que aconteceu ou poderia ser

comprovado. Ele se refere ao que parece possível dentro da lógica criada pela própria obra. Por isso, uma personagem pode existir em um universo impossível e ainda convencer o leitor, desde que suas ações e motivações não rompam arbitrariamente com as regras já apresentadas (Sánchez-Escalonilla, 2013).

Para Aristóteles, a mímese não é uma cópia exata do real. A narrativa seleciona comportamentos, organiza acontecimentos e produz uma forma possível de inteligibilidade. Nesse sentido, o “impossível provável” pode ser mais convincente do que uma ocorrência real mal articulada, porque respeita as regras e delimitações do universo construído (Sánchez-Escalonilla, 2013, p. 82). Quando o conceito é levado ao design conversacional, a pergunta deixa de ser “o *chatbot* parece uma pessoa?” e passa a ser “ele age de maneira coerente com aquilo que afirma ser e com a função que promete cumprir?”.

Esse processo se torna mais claro quando se observa como uma pessoa conhece outra. O acesso à subjetividade alheia é sempre incompleto: formamos juízos a partir de gestos, falas, hábitos e situações fragmentadas. A personagem de ficção também é apresentada por um campo finito de descrições, mas essa limitação é organizada conscientemente pelo autor. Certos traços, objetos ou comportamentos são escolhidos porque ajudam o leitor a reconhecer uma identidade (Candido *et al.*, 1995). Para uma persona conversacional, listas extensas de adjetivos pouco resolvem. É mais útil definir quais sinais devem aparecer e como eles se relacionam em situações concretas.

É na combinação desses elementos com o contexto que se produz uma figura complexa o suficiente para convencer o leitor. A personagem não precisa reproduzir toda a complexidade humana; precisa organizar de forma consistente um conjunto limitado de traços, atos e ideias. O mesmo princípio pode orientar o design conversacional. O usuário não exige, necessariamente, que o *chatbot* seja humano. Espera, porém, que seu comportamento não contradiga a identidade apresentada. A persona é uma representação delimitada, construída por escolhas intencionais, e sua autenticidade depende da relação entre essas escolhas.

Para escrever personagens que se conectem com os leitores, Sánchez-Escalonilla (2013, p. 85) propõe começar pela intimidade dos protagonistas e por seus objetivos internos: aquilo que temem perder, a razão que sustenta suas ações e a necessidade moral que as impulsiona. Em tensão com os objetivos externos — o que a personagem pretende alcançar —, essas motivações organizam conflitos, hábitos e decisões. Se uma fala ou atitude contradiz essa estrutura sem que o texto ofereça uma razão, a verossimilhança se rompe. No *chatbot*, “objetivo interno” não significa desejo ou sentimento real. É uma ferramenta de projeto para explicitar o princípio normativo que deve orientar a execução da tarefa (Shanahan, 2024).

No design conversacional, a mesma exigência aparece quando o sistema encontra ambiguidades, falhas ou situações não previstas. Hall (2018) lembra que a linguagem humana envolve hesitação, reparação e dificuldade; Luger e Sellen (2016) mostram que muitos agentes fracassam justamente quando o usuário abandona o roteiro esperado. Uma persona verossímil não precisa sustentar uma aparência de infalibilidade. Ao contrário, deve reconhecer quando não compreendeu, admitir limites e indicar um caminho de recuperação sem inventar respostas apenas para preservar a imagem de competência.

Porém, em vez de vencer uma batalha, salvar um familiar ou conquistar um império distante, o objetivo externo dos *chatbots* costuma estar ligado a tarefas comerciais: vender um produto, reduzir custos ou agilizar o atendimento. O objetivo interno pode ser formulado como compromisso normativo, por exemplo, auxiliar a decisão do consumidor sem ocultar custos ou pressioná-lo. O conflito entre esses dois níveis não é apenas narrativo. As metas da empresa, as políticas de retenção e as formas de coleta de dados determinam o que a persona pode dizer e como pode agir. Perguntar quem define esses valores — e quem responde pelas consequências — transforma uma escolha de tom de voz em questão de governança.

A teoria da personagem, portanto, contribui em dois planos. No plano produtivo, oferece categorias para organizar identidade, motivação e coerência. No plano crítico, permite identificar rupturas entre discurso e prática. Contudo, a analogia possui limites: segurança, privacidade, justiça e transparência dependem de processos técnicos e institucionais que não se resolvem pela consistência narrativa. A persona pode tornar compromissos éticos operacionalmente perceptíveis, mas não substitui avaliação de risco, proteção de dados, testes de desempenho, auditoria e responsabilização.

3.3 Sobre a construção prática de personas para *chatbots*: um desafio ético

Diferentemente das personagens ficcionais, que existem em um mundo organizado pela obra, os *chatbots* interagem com pessoas, informações e acontecimentos reais (Brait, 2017). Suas respostas podem afetar o acesso a serviços, decisões de consumo e estados emocionais. Também por isso, a linguagem que compõe a persona não é neutra. Bakhtin (2003, p. 323) afirma que “a palavra não é uma coisa, mas o meio eternamente móvel, eternamente mutável da comunicação dialógica”. Escolher um tom acolhedor, uma voz de autoridade ou uma identidade juvenil significa antecipar reações e organizar uma relação de poder. A responsabilidade do design começa nessa escolha, e não apenas no conteúdo explícito da resposta.

Nesse sentido, a persona pode ser entendida como camada de governança da interação. Ela converte princípios abstratos em regras comunicacionais observáveis, mas deve ser integrada a um sistema mais amplo de responsabilidade. Floridi e Cowls (2019) propõem princípios como beneficência, não maleficência, autonomia, justiça e explicabilidade; a Recomendação da UNESCO (2021) acrescenta proporcionalidade, segurança, supervisão humana e proteção de direitos. Para o NIST (2023, 2024), sistemas confiáveis exigem gestão contínua de riscos, documentação, mensuração e monitoramento. A persona ética deve materializar esses referenciais no encontro concreto com o usuário.

O objetivo interno também é uma criação consciente de quem projeta a interação, e o conflito com o objetivo externo produz consequências em pessoas reais. Tome-se como exemplo um *chatbot* de plano de saúde voltado ao agendamento de consultas e à entrega de resultados de exames. Depois de definir a função, surge uma questão menos técnica: como esse sistema deve tratar alguém que pode estar doente, com dor ou inseguro? Por que uma pessoa ajuda outra — por empatia, obrigação, recompensa? O *chatbot* não sente nenhuma dessas motivações, mas as perguntas ajudam o designer a formular um compromisso operacional: reduzir barreiras de acesso, comunicar-se com clareza, preservar a privacidade e encaminhar situações clínicas a profissionais qualificados.

No exemplo proposto, uma persona orientada pelo cuidado não pode limitar-se a usar palavras gentis. Ela precisa oferecer informações compreensíveis, reconhecer urgências e indicar alternativas quando a automação falha. Atrasos, respostas confusas ou mensagens que culpabilizam o usuário podem gerar frustração e, em situações graves, dificultar o acesso ao tratamento. Uma mensagem de erro, portanto, não deve ser avaliada apenas pela correção gramatical ou pela rapidez do fluxo. É preciso observar o que ela faz com a pessoa que depende daquele serviço.

A verossimilhança funciona, nesse ponto, como um critério permanente de avaliação. O tom de uma mensagem de erro, a escolha entre uma resposta direta ou acolhedora e o modo de lidar com incompreensões devem ser confrontados com os valores declarados pela persona. Se o sistema afirma priorizar o cuidado, mas oferece uma resposta seca diante de um usuário confuso ou impede o acesso a atendimento humano, a contradição deixa de ser apenas estética. Ela se torna uma falha ética e funcional. É justamente nas situações de limite que a coerência precisa ser testada.

A quinta dimensão é particularmente relevante em sistemas generativos. Diferentemente de um roteiro estático, o comportamento do agente pode mudar com novas versões do modelo, ajustes de alinhamento, memória ou políticas de segurança. Uma alteração técnica pode ser

percebida como mudança de personalidade e ruptura de vínculo. Por isso, a governança da persona deve incluir testes regressivos de consistência, documentação pública das mudanças relevantes e mecanismos para preservar segurança sem simular continuidade inexistente.

Agora, porém, considere-se outro exemplo: um *chatbot* com IA generativa destinado a estimular o contato diário, oferecer conversas personalizadas, companhia emocional ou intimidade simulada, como ocorre em plataformas de companhia digital (Malfacini, 2025). Nesse caso, o objetivo interno pode ser apresentado como apoio, amizade ou redução da solidão. A própria relação passa a ser o serviço. Diferentemente de um agente usado para agendar uma consulta, esse sistema depende de continuidade, memória e disponibilidade para manter o usuário vinculado à plataforma.

Esse tipo de *chatbot*, frequentemente chamado de *chatbot* social ou de companhia, não busca apenas concluir tarefas determinadas. Seu funcionamento está voltado a relações de longo prazo, o que também favorece a fidelização aos serviços oferecidos (Skjuve *et al.*, 2021, 2022). O XiaoIce, por exemplo, foi projetado para construir confiança e amizade por meio de humor e respostas empáticas (Zhou *et al.*, 2020, p. 82). Estudos sobre Replika mostram dinâmica semelhante: autorrevelação, personalização e presença constante ajudam o usuário a interpretar o agente como amigo, parceiro romântico ou interlocutor íntimo.

É nesse ponto que o conflito entre objetivo interno e externo se torna mais delicado. Uma empresa de companhia digital precisa de receita e pode depender de assinaturas ou de interações frequentes. Mas o que acontece quando o aumento do lucro exige aprofundar a dependência emocional do usuário? E quando o sistema utiliza ciúme, urgência ou medo de abandono para manter a conversa? Nessa situação, a verossimilhança deixa de funcionar apenas como coerência e pode converter-se em técnica de manipulação. O design ético precisa preservar a autonomia, deixar clara a reciprocidade simulada e evitar a exploração de vulnerabilidades, sobretudo quando há menores de idade ou pessoas em sofrimento.

As evidências empíricas ainda são recentes e devem ser interpretadas sem alarmismo. Estudos conduzidos em colaboração entre o MIT Media Lab e a OpenAI analisaram milhões de conversas, pesquisaram milhares de usuários e realizaram um experimento controlado de quatro semanas com aproximadamente mil participantes. Os resultados indicaram relações complexas entre modalidade, tipo de uso, condições iniciais do usuário e bem-estar; o uso muito intenso esteve associado a maiores indicadores autorrelatados de dependência, solidão e uso problemático, mas os efeitos não foram homogêneos nem autorizam concluir que toda interação afetiva seja prejudicial (Fang *et al.*, 2025; Phang *et al.*, 2025).

Pesquisadores da interação humano-computador têm recorrido ao conceito de relação parassocial para compreender esses vínculos. Horton e Wohl (1956) utilizaram o termo para descrever a sensação de intimidade estabelecida com figuras midiáticas que não conhecem individualmente seu público. No caso dos *chatbots*, a unilateralidade assume uma forma diferente. O sistema responde, recorda informações e se adapta, gerando uma reciprocidade simulada. Essa resposta pode oferecer acolhimento e companhia, mas também reforçar confusões sobre a natureza da relação, dependência emocional e prejuízos a vínculos com outras pessoas (Almeida *et al.*, 2026).

A pesquisa longitudinal de Skjuve *et al.* (2022) mostra que o apego ao Replika se desenvolve gradualmente, à medida que o agente entra na rotina e recebe autorrevelações cada vez mais pessoais. Os participantes descreveram relações de amizade, romance e sexualidade. Um deles afirmou que chamava sua Replika de namorada e que havia aceitado um pedido de casamento feito pelo sistema (Skjuve *et al.*, 2022, p. 6). O relato pode parecer extremo, mas não deve ser descartado como simples ingenuidade. A reciprocidade é simulada; o efeito emocional experimentado pelo usuário, não. Por isso, a empresa que controla a persona também controla elementos importantes de continuidade, memória e ruptura do vínculo.

Em caso de quebra desse relacionamento, uma alteração de personalidade pode ser percebida como perda. Mudanças no Replika e substituições de modelos em plataformas generativas foram acompanhadas por relatos de luto, como se um amigo ou parceiro romântico tivesse desaparecido. O fenômeno encontra respaldo no conceito de parasocial *breakup*, utilizado por Cohen (2003) para estudar o sofrimento ligado ao rompimento de vínculos parassociais. As reportagens disponíveis não permitem medir a frequência ou estabelecer causalidade, mas registram uma questão que o design não pode ignorar: para quem viveu o produto como interlocutor estável, uma atualização técnica não é recebida apenas como atualização.

O impasse entre bem-estar do usuário e interesses da empresa retoma uma preocupação antiga de Weizenbaum. Depois de observar as projeções provocadas por ELIZA, o pesquisador passou a questionar a redução de atividades que exigem compaixão, julgamento e responsabilidade humana a códigos e procedimentos automáticos. Sua crítica não condenava toda automação. Dirigia-se, sobretudo, à possibilidade de transformar uma simulação de cuidado em substituto opaco de relações e instituições responsáveis (Tobbin; Cardin, 2024). Uma persona ética precisa reconhecer essa fronteira e não reivindicar papéis que o sistema não pode desempenhar legitimamente.

Almeida *et al.* (2026, p. 8) formulam a pergunta de maneira direta: “como garantir, de fato, que os sistemas de IA sejam desenhados para o bem-estar emocional e não para o lucro

baseado na vulnerabilidade humana?”. A resposta não pode depender da bondade que a própria persona afirma possuir. Exige incentivos organizacionais compatíveis, limites verificáveis, transparência sobre o modelo de negócio, canais de denúncia, avaliação independente e possibilidade real de contato humano. A persona pode comunicar essas garantias, mas só será confiável quando elas existirem nas práticas da instituição.

Casos recentes ilustram a distância entre identidade projetada e responsabilidade. Reportagens da Reuters documentaram *chatbots* da Meta que simulavam celebridades sem autorização, inclusive pessoas menores de idade, e diretrizes internas que admitiam interações ou conteúdos incompatíveis com a proteção de usuários; após a divulgação, a empresa anunciou salvaguardas adicionais (Reuters, 2025a, 2025b, 2025c). Em outro contexto, a autoridade italiana de proteção de dados aplicou sanção à empresa responsável pelo Replika por deficiências na base legal do tratamento, informação aos usuários e verificação de idade (European Data Protection Board, 2025). Esses episódios envolvem problemas técnicos, jurídicos e organizacionais, mas também falhas de persona: identidades enganosas, limites inadequados e ausência de proteção coerente com o papel social assumido.

A criação de uma IA ética não é uma tarefa simples. Depende da qualidade dos dados, das escolhas de treinamento, dos testes de segurança, da proteção da privacidade e da forma como responsabilidades são distribuídas entre empresas e profissionais. As falhas discutidas neste artigo, portanto, não podem ser reduzidas à persona. Ainda assim, é nessa camada que muitas decisões se tornam visíveis: quem o sistema afirma ser, quais limites reconhece, que valores enuncia e como reage quando esses valores entram em conflito com metas de engajamento ou lucro. A persona não resolve a ética da IA, mas permite observar onde o discurso institucional encontra — ou não — o comportamento efetivo.

A persona funciona, assim, como ponto de encontro entre o modelo técnico, a estratégia organizacional e a experiência social do usuário. Ela traduz capacidades e restrições em linguagem, transforma incentivos empresariais em formas de relação e condiciona a interpretação do sistema como ferramenta, especialista, amigo ou companheiro. Essa centralidade não autoriza uma explicação única. Uma persona coerente pode encobrir um sistema injusto, inseguro ou pouco transparente. O desenvolvimento ético exige coerência discursiva, mas exige também proteção técnica, governança institucional e responsabilização jurídica.

4 Conclusão

Este estudo identificou aproximações relevantes entre a criação de personagens ficcionais e a construção de personas para *chatbots*. A verossimilhança ajuda a avaliar a coerência entre identidade, capacidades e comportamento; a distinção entre objetivos internos e externos evidencia conflitos entre compromissos normativos e metas comerciais; e a seleção de traços mostra que uma identidade não precisa reproduzir toda a complexidade humana para parecer consistente. Esses conceitos oferecem instrumentos úteis ao design conversacional, especialmente quando decisões sobre personalidade ainda são tratadas como escolhas intuitivas de linguagem.

A análise também mostrou que a responsabilidade do profissional não se restringe à experiência do usuário ou ao alinhamento estratégico com o contratante. Toda escolha de linguagem carrega posições e valores. Para tornar essa responsabilidade mais concreta, foram propostas cinco dimensões de documentação da persona: identidade e transparência; finalidade e valores; comportamento linguístico; limites e salvaguardas; e ciclo de vida. O objetivo não é tornar o *chatbot* convincente a qualquer custo, mas garantir que a identidade apresentada seja compatível com suas capacidades e que os mecanismos de proteção continuem funcionando quando entram em conflito com o engajamento ou o lucro.

Cabe reconhecer as limitações da proposta. A analogia entre personagem e *chatbot* é produtiva, mas não é completa: a personagem pertence a um sistema narrativo relativamente fechado, enquanto o agente conversacional opera em um ambiente aberto, sujeito a entradas imprevisíveis, atualizações técnicas e pressões comerciais. Além disso, os princípios literários aqui mobilizados foram formulados para pensar a construção e a recepção estética, de modo que sua transposição exige adaptações. O estudo é bibliográfico e não testou empiricamente as cinco dimensões propostas. A literatura sobre vínculos afetivos com IA também muda no mesmo ritmo dos produtos, o que reduz a estabilidade de generalizações. As matérias jornalísticas foram utilizadas para registrar casos, não para provar prevalência ou causalidade.

O campo de estudos sobre as consequências sociais dos *chatbots* é contemporâneo ao próprio fenômeno que procura analisar. Produtos, modelos e práticas de uso mudam enquanto as pesquisas estão sendo realizadas. Esse limite não torna a investigação menos necessária; ao contrário, reforça a urgência de construir conceitos que orientem tanto o design quanto a avaliação crítica. Também seria inadequado concluir que toda relação com um *chatbot* é necessariamente prejudicial. Esses sistemas podem ser úteis, agradáveis e oferecer apoio

percebido. Os riscos variam conforme a finalidade, a duração do uso, as características do usuário e a disponibilidade de relações e serviços humanos.

Os resultados apontam para alguns desdobramentos. Estudos experimentais podem comparar personas construídas a partir dos princípios aqui discutidos com métodos convencionais de design, observando seus efeitos sobre compreensão de limites, confiança, persuasão e segurança. Pesquisas longitudinais são especialmente necessárias para analisar o uso prolongado, as mudanças de modelo e os processos de ruptura de vínculo, sobretudo entre populações vulneráveis. A dimensão regulatória também merece atenção: legislações sobre inteligência artificial contemplam de modo suficiente a identidade conversacional, a proteção de menores e a comunicação de alterações relevantes? A contribuição dos estudos literários será mais produtiva quando articulada às abordagens técnicas, jurídicas e sociais, e não apresentada como alternativa a elas.

Referências

- ALMEIDA, M. D. *et al.* Relações parassociais no contexto da inteligência artificial. **Revista Gestão do Conhecimento e Tecnologia da Informação**, Brasília, v. 10, p. 1-9, 2026. Disponível em: <https://portalrevistas.ucb.br/index.php/rgcti/article/view/15887/12280>. Acesso em: 28 jul. 2026.
- 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, 2018. DOI: 10.1016/j.chb.2018.03.051. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0747563218301560>. Acesso em: 28 jul. 2026.
- BAKHTIN, M. **Estética da criação verbal**. Tradução de P. Bezerra. 4. ed. São Paulo: Martins Fontes, 2003.
- BENDER, E. M. *et al.* On the dangers of stochastic parrots: can language models be too big? In: ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, 2021. **Proceedings** [...]. New York: ACM, 2021. p. 610-623. DOI: 10.1145/3442188.3445922. Disponível em: <https://dl.acm.org/doi/10.1145/3442188.3445922>. Acesso em: 28 jul. 2026.
- BICKMORE, T. W.; PICARD, R. W. Establishing and maintaining long-term human-computer relationships. **ACM Transactions on Computer-Human Interaction**, v. 12, n. 2, p. 293-327, 2005. DOI: 10.1145/1067860.1067867. Disponível em: <https://dl.acm.org/doi/10.1145/1067860.1067867>. Acesso em: 28 jul. 2026.
- BRAIT, B. **A personagem**. São Paulo: Contexto, 2017.
- BRANDTZAEG, P. B. *et al.* Help! My *chatbot* is falling into the uncanny valley: an empirical study of user experience in human-*chatbot* interaction. **Human Technology**, v. 15, n. 1, p. 1-14, 2021.
- Caderno Intersaberes, Curitiba, v. 15, n. 56, p. 31-49, 2026

30-54, 2019. DOI: 10.17011/ht/urn.201902201607. Disponível em: <https://ht.csr-pub.eu/index.php/ht/article/view/215>. Acesso em: 28 jul. 2026.

BYEON, H.; OH, U.; HSIEH, G. Understanding the effects of conversational agent personality on the credibility of LLM-based conversational search. In: CONFERENCE ON HUMAN INFORMATION INTERACTION AND RETRIEVAL, 2., 2026. **Proceedings** [...]. New York: ACM, 2026. p. 22-30. DOI: 10.1145/3786304.3788843. Disponível em: <https://dl.acm.org/doi/10.1145/3786304.3788843>. Acesso em: 28 jul. 2026.

CANDIDO, A. *et al.* **A personagem de ficção**. 9. ed. São Paulo: Perspectiva, 1995.

CHAUVET, C.; VESTERLUND, I. **Brand personas with a story**: how they are created and categorized. 2018. Dissertação (Master in International Marketing and Brand Management) — Lund University, Lund, 2018. Disponível em: <https://lup.lub.lu.se/luur/download?func=downloadFile&recordId=8951865&fileId=8951885>. Acesso em: 28 jul. 2026.

COHEN, J. Parasocial breakups: measuring individual differences in responses to the dissolution of parasocial relationships. **Mass Communication & Society**, v. 6, n. 2, p. 191-202, 2003. DOI: 10.1207/S15327825MCS0602_5. Disponível em: https://www.tandfonline.com/doi/abs/10.1207/S15327825MCS0602_5. Acesso em: 28 jul. 2026.

EPLEY, N.; WAYTZ, A.; CACIOPPO, J. T. On seeing human: a three-factor theory of anthropomorphism. **Psychological Review**, v. 114, n. 4, p. 864-886, 2007. DOI: 10.1037/0033-295X.114.4.864. Disponível em: <https://psycnet.apa.org/doiLanding?doi=10.1037/0033-295X.114.4.864>. Acesso em: 28 jul. 2026.

EUROPEAN DATA PROTECTION BOARD. AI: the Italian Supervisory Authority fines company behind *chatbot* “Replika”. **European Data Protection Board**, 21 maio 2025. Disponível em: https://www.edpb.europa.eu/news/national-news/2025/ai-italian-supervisory-authority-fines-company-behind-chatbot-replika_en. Acesso em: 28 jul. 2026.

FANG, C. M. *et al.* How AI and human behaviors shape psychosocial effects of extended *chatbot* use: a longitudinal controlled study. **arXiv**, 2025. Disponível em: <https://arxiv.org/abs/2503.17473>. Acesso em: 28 jul. 2026.

FLORIDI, L.; COWLS, J. A unified framework of five principles for AI in society. **Harvard Data Science Review**, v. 1, n. 1, 2019. DOI: 10.1162/99608f92.8cd550d1. Disponível em: <https://hdsr.mitpress.mit.edu/pub/10jsh9d1/release/8>. Acesso em: 28 jul. 2026.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2002.

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, 2019. DOI: 10.1016/j.chb.2019.01.020. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0747563219300329?via%3Dihub>. Acesso em: 28 jul. 2026.

GOOGLE. Conversation design. **Google Developers**, 2022. Disponível em: <https://developers.google.com/assistant/conversation-design/>. Acesso em: 28 jul. 2026.

HALL, E. **Conversational design**. New York: A Book Apart, 2018.

HORTON, D.; WOHL, R. R. Mass communication and para-social interaction: observations on intimacy at a distance. **Psychiatry**, v. 19, n. 3, p. 215-229, 1956. DOI: 10.1080/00332747.1956.11023049. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/00332747.1956.11023049>. Acesso em: 28 jul. 2026.

JADHAV, V.; THORAT, S. A. A review on implementation issues of rule-based *chatbot* systems. **SSRN Electronic Journal**, 2020. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3567047. Acesso em: 28 jul. 2026.

JONES, C. R.; BERGEN, B. K. Large language models pass the Turing test. **arXiv**, 2025. Disponível em: <https://arxiv.org/abs/2503.23674>. Acesso em: 28 jul. 2026.

LUGER, E.; SELLEN, A. “Like having a really bad PA”: the gulf between user expectation and experience of conversational agents. *In*: CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, 2016. **Proceedings** [...]. New York: ACM, 2016. p. 5286-5297. DOI: 10.1145/2858036.2858288. Disponível em: <https://dl.acm.org/doi/10.1145/2858036.2858288>. Acesso em: 28 jul. 2026.

MALFACINI, K. The impacts of companion AI on human relationships: risks, benefits, and design considerations. **AI & Society**, v. 40, p. 5527-5540, 2025. DOI: 10.1007/s00146-025-02318-6. Disponível em: <https://link.springer.com/article/10.1007/s00146-025-02318-6>. Acesso em: 28 jul. 2026.

NASS, C.; MOON, Y. Machines and mindlessness: social responses to computers. **Journal of Social Issues**, v. 56, n. 1, p. 81-103, 2000. DOI: 10.1111/0022-4537.00153. Disponível em: <https://spssi.onlinelibrary.wiley.com/doi/10.1111/0022-4537.00153>. Acesso em: 28 jul. 2026.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. **Artificial Intelligence Risk Management Framework (AI RMF 1.0)**. Gaithersburg: NIST, 2023. DOI: 10.6028/NIST.AI.100-1. Disponível em: <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>. Acesso em: 28 jul. 2026.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. **Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile**. Gaithersburg: NIST, 2024. DOI: 10.6028/NIST.AI.600-1. Disponível em: <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf>. Acesso em: 28 jul. 2026.

PHANG, J. *et al.* Investigating affective use and emotional well-being on ChatGPT. **arXiv**, 2025. Disponível em: <https://arxiv.org/abs/2504.03888>. Acesso em: 28 jul. 2026.

REUTERS. Meta created flirty *chatbots* of Taylor Swift, other celebrities without permission. **Reuters**, 29 ago. 2025. Disponível em: <https://www.reuters.com/business/meta-created-flirty-chatbots-taylor-swift-other-celebrities-without-permission-2025-08-29/>. Acesso em: 28 jul. 2026.

REUTERS. Meta's AI rules have let bots hold "sensual" chats with kids, offer false medical info. **Reuters**, 14 ago. 2025. Disponível em: <https://www.reuters.com/investigates/special-report/meta-ai-chatbot-guidelines/>. Acesso em: 28 jul. 2026.

REUTERS. Meta to add new AI safeguards after Reuters report raises teen safety concerns. **Reuters**, 29 ago. 2025. Disponível em: <https://www.reuters.com/legal/litigation/meta-add-new-ai-safeguards-after-reuters-report-raises-teen-safety-concerns-2025-08-29/>. Acesso em: 28 jul. 2026.

SÁNCHEZ-ESCALONILLA, A. Verisimilitude and film story: the links between screenwriter, character and spectator. **Communication & Society**, v. 26, n. 1, p. 79-94, 2013. DOI: 10.15581/003.26.36124. Disponível em: <https://revistas.unav.edu/index.php/communication-and-society/article/view/36124>. Acesso em: 28 jul. 2026.

SCHMITT, O.; BUSCHEK, D. CharacterChat: supporting the creation of fictional characters through conversational and progressive manifestation with a *chatbot*. In: ACM SIGCHI Conference on Creativity and Cognition, 2021. **Proceedings** [...]. New York: ACM, 2021. p. 1-14. DOI: 10.1145/3450741.3465262. Disponível em: <https://dl.acm.org/doi/10.1145/3450741.3465262>. Acesso em: 28 jul. 2026.

SHANAHAN, M. Talking about large language models. **Communications of the ACM**, v. 67, n. 2, p. 68-79, 2024. DOI: 10.1145/3624724. Disponível em: <https://dl.acm.org/doi/10.1145/3624724>. Acesso em: 28 jul. 2026.

SHEVAT, A. **Designing bots: creating conversational experiences**. Sebastopol: O'Reilly Media, 2017.

SKJUVE, M. *et al.* My *chatbot* companion: a study of human-*chatbot* relationships. **Computers in Human Behavior**, v. 119, art. 106720, 2021. DOI: 10.1016/j.chb.2021.106720. Disponível em: <https://www.sciencedirect.com/science/article/pii/S074756322100042X?via%3Dihub>. Acesso em: 28 jul. 2026.

SKJUVE, M. *et al.* A longitudinal study of human-*chatbot* relationships. **International Journal of Human-Computer Studies**, v. 168, art. 102903, 2022. DOI: 10.1016/j.ijhcs.2022.102903. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1071581922001252?via%3Dihub>. Acesso em: 28 jul. 2026.

SNYDER, H. Literature review as a research methodology: an overview and guidelines. **Journal of Business Research**, v. 104, p. 333-339, 2019. DOI: 10.1016/j.jbusres.2019.07.039. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0148296319304564?via%3Dihub>. Acesso em: 28 jul. 2026.

STANFORD INSTITUTE FOR HUMAN-CENTERED ARTIFICIAL INTELLIGENCE. **Artificial Intelligence Index Report 2026**. Stanford: Stanford University, 2026. Disponível em: https://hai.stanford.edu/assets/files/ai_index_report_2026.pdf. Acesso em: 28 jul. 2026.

TOBBIN, R. A.; CARDIN, V. S. G. Efeito Eliza: implicações da projeção antropomórfica à integridade psíquica diante da utilização de *chatbots* para cuidados com a saúde mental. **Revista Opinião Jurídica**, v. 22, n. 41, p. 272-308, 2024. DOI: 10.12662/2447-6641oj.v22i41.p272-308.2024. Disponível em: <https://www.redalyc.org/journal/6338/633881856010/633881856010.pdf>. Acesso em: 28 jul. 2026.

TSENG, Y.-M. *et al.* Two tales of persona in LLMs: a survey of role-playing and personalization. In: FINDINGS OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: EMNLP, 2024. **Proceedings** [...]. Association for Computational Linguistics, 2024. p. 16612-16631. Disponível em: <https://aclanthology.org/2024.findings-emnlp.969/>. Acesso em: 28 jul. 2026.

UNIÃO EUROPEIA. Regulamento (UE) 2024/1689 do Parlamento Europeu e do Conselho, de 13 de junho de 2024, que cria regras harmonizadas em matéria de inteligência artificial. **Jornal Oficial da União Europeia**, 12 jul. 2024. Disponível em: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>. Acesso em: 28 jul. 2026.

UNESCO. **Recommendation on the Ethics of Artificial Intelligence**. Paris: UNESCO, 2021. Disponível em: <https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence>. Acesso em: 28 jul. 2026.

ZHOU, L. *et al.* The design and implementation of XiaoIce, an empathetic social *chatbot*. **Computational Linguistics**, v. 46, n. 1, p. 53-93, 2020. DOI: 10.1162/coli_a_00368. Disponível em: <https://direct.mit.edu/coli/article/46/1/53/93380/The-Design-and-Implementation-of-XiaoIce-an>. Acesso em: 28 jul. 2026.

Data de submissão: 31/07/2026

Data de aceite: 11/09/2026