

# Erros, falhas e perturbações digitais em alucinações das IA generativas: Tipologia, premissas e epistemologia da comunicação

## *Digital errors, failures, and disruptions in generative AI hallucinations: Communication typology, premises, and epistemology*

■ ANDRÉ LUIZ MARTINS LEMOS<sup>a</sup>

Universidade Federal da Bahia, Faculdade de Comunicação, Programa de Pós-Graduação em Comunicação e Cultura Contemporâneas. Salvador – BA, Brasil

### RESUMO

Neste artigo, identifica-se como erros, falhas e perturbações digitais podem ser analisados a partir de três premissas, relacionando-os com o exemplo da atual alucinação em sistema de IAG, como o ChatGPT. Eles revelam uma dimensão escondida dos objetos digitais. Por serem mais concretos, os objetos digitais geram uma maior indefinição das origens e consequências de eventos disruptivos. Nesses momentos, pode-se vislumbrar agenciamentos coletivos em torno da cultura digital. A proposta é que tais erros, falhas e perturbações sejam entendidos não como positivos ou negativos, mas como uma forma de apontar direcionamentos para a pesquisa, indicando o *locus* para uma abordagem qualitativa. Conclui-se que os erros não são apenas disruptivos, ou oportunidades para gerar inovação, mas eventos que permitem entender as formas da comunicação e as ações das mídias digitais.

**Palavras-chave:** Erros digitais, alucinação, ChatGPT, IA, neomaterialismo

<sup>a</sup> Professor Titular da Faculdade de Comunicação da Universidade Federal da Bahia (UFBA). Professor permanente do Programa de Pós-Graduação em Comunicação e Cultura Contemporâneas (Facom/UFBA). Pesquisador IA do CNPQ. Coordenador do Lab404. Orcid: <http://orcid.org/0000-0001-9291-6494>. E-mail: [almlemos@gmail.com](mailto:almlemos@gmail.com)

### ABSTRACT

This study finds how digital errors, failures, and disruptions may be analyzed based on three premises, relating them to contemporary hallucination in generative AI systems, such as ChatGPT. These factors show a hidden dimension of digital objects. Because digital objects are more concrete, they generate greater uncertainty about the origins and consequences of disruptive events, moments which enable a glimpse of the collective agencies around the digital culture. This study proposes that errors, failures, and disruptions (positive or negative) point toward directions for research and indicate a locus for a qualitative approach. It is concluded that errors are not only

DOI: <http://dx.doi.org/10.11606/issn.1982-8160.v18i1p75-91>



# D

## Erros, falhas e perturbações digitais em alucinações das IA generativas

disruptive (as an opportunity to generate innovation) but also events that enable us to understand the forms of communication and actions of digital media.

**Keywords:** Digital errors, hallucination, ChatGPT, AI, new materialism

OS DEBATES SOBRE o ChatGPT circulam sob o signo do erro, da perturbação e da falha. Ele tem chamado a atenção por seu caráter disruptivo, seja no uso *correto*, ameaçando postos de trabalho e a própria existência da humanidade, seja quando alucina, gerando informações falsas. Ao destacar as benesses da inteligência artificial generativa (IAG), tipo de inteligência artificial que tem a capacidade de gerar conteúdo (dados, texto, imagens, sons) com base em algoritmos de aprendizado, redes neurais e treinamento em um grande conjunto de dados, o debate vem sempre permeado pelos seus erros, falhas ou potenciais perturbações.

Muitos estudos apontam como erros, falhas e perturbações são momentos privilegiados para pensar o social e a cultura digital, em particular (Alexander, 2017; Appadurai & Alexander, 2020; Barker & Korolkova, 2022; Bellinger, 2016; Korolkova & Bowes, 2020; Nunes, 2011; Parikka & Sampson, 2009; Rettberg, 2022). Esses momentos ajudam a entender a vida dos objetos, os seus agenciamentos, as formas de concretização e apontam para uma dimensão metodológica e epistemológica do próprio erro.

Em um primeiro momento, definem-se erros, falhas e perturbações na cultura digital. Em seguida, descreve-se o ChatGPT e o conceito de “alucinação algorítmica”. Através de um diálogo com essa IAG, apontamos concretamente como ela alucina. A terceira parte do artigo expõe as premissas de pesquisas sobre erros digitais e as relaciona com a alucinação do ChatGPT. A última parte aborda a discussão dos erros na cultura digital como um método para vislumbrar os agenciamentos em jogo. Conclui-se que os erros não são apenas disruptivos, ou oportunidade para gerar inovação, mas eventos que permitem entender as formas da comunicação e as ações das mídias digitais.

### DEFINIÇÕES ERROS, FALHAS E PERTURBAÇÕES

Para entender o problema do erro (colocado aqui ainda de forma genérica) na cultura digital, proponho pensar em três categorias: erros, falhas e perturbações. Há uma certa confusão entre os termos. Erros podem gerar outros erros, falhas ou perturbações. Por exemplo, uma citação errada de uma IAG pode levar a erro de argumento em um artigo científico (erro). Erros em notícias (factuais) podem alimentar desinformação (perturbação). Falhas em infraestruturas (elétrica ou

de conexão, por exemplo) podem ocasionar erros lógicos em sistemas e efeitos perturbadores, como incapacidade de acessar aplicativos de bancos ou redes sociais. Toda mutação tecnológica gera erros, falhas e perturbações, permitindo entender o lugar de sua produção, o contexto histórico e os arranjos sociais que lhe dão guarida. Para um melhor diagnóstico dos efeitos disruptivos da cultura digital, proponho diferenciar erros, falhas e perturbações.

Erros são problemas lógicos, ou de princípios, que desviam do resultado pretendido. A definição está vinculada ao sentido original da palavra, errância, desvio. Erros são eventos desviantes gerados por problemas de lógica, princípio, funcionamento interno. Como diz Nunes, o “erro marca um desvio de um resultado predeterminado”<sup>1</sup> (2011, p. 7). Portanto, eles não são fenômenos externos, embora possam ser gerados por eles (que definiremos como falhas), mas parte integral do funcionamento dos sistemas cibernético-informacionais. Um erro pode produzir falhas e perturbações, mas estas nem sempre são geradas por ele. Um erro lógico pode endereçar a um usuário uma informação equivocada, por exemplo, gerando falhas e perturbações no dispositivo ou no sistema como um todo. O erro de endereçamento (DNS) no Facebook em 2021 gerou falhas e perturbações pois o trabalho por WhatsApp foi interrompido em algumas empresas e pessoas ficaram sem saber como agir sem acesso às plataformas da Meta<sup>2</sup>.

<sup>1</sup>No original: “Error marks a deviation from a predetermined outcome”.

<sup>2</sup>Ver Taylor (2021).

Falhas são problemas gerados por eventos externos, podendo gerar erros e perturbações. Elas podem ser ocasionadas por erros lógicos ou por fenômenos externos, como problemas de infraestrutura (pane elétrica, corte de cabos de conexão, acesso a servidores etc.), acidentes ou desastres naturais, uso equivocado de sistemas, entre outros. A análise de Bennett (2005) sobre o *blackout* elétrico nos Estados Unidos em 2003 mostra bem como as falhas geraram erros e perturbações, revelando agenciamentos múltiplos como causa e consequência do evento.

Já as perturbações são eventos disruptivos, na maioria das vezes causados pelo uso em sintonia com a lógica e gramática dos sistemas ou plataformas (Dourish, 2017; Gillespie, 2010), podendo ou não serem originadas por erros ou falhas. Essas perturbações podem ser vistas como anomalias (Parikka & Sampson, 2009) que tensionam questões éticas, morais e políticas de acordo com o contexto histórico. Na sociologia, perturbações afetam um coletivo (sistema social) e desencadeiam mudanças (teoria da mudança social) (Arendt, 1998; Giddens, 1990; Marx & Engels, 2011; Weber, 2001). Elas podem ter diversas origens (tecnológicas, científicas, conflitos políticos, ondas econômicas, desastres naturais, guerras, migrações etc.).

# D

## Erros, falhas e perturbações digitais em alucinações das IA generativas

As perturbações digitais (tecnológicas) são ocasionadas pelo uso corriqueiro dos dispositivos (celulares, computadores e internet mudaram e continuam a mudar a sociedade) e ameaçam valores e conquistas sociais (enviesamento de gênero e raça, problema ambiental, trabalho na *gig economy*, regime econômico de controle e vigilância, entre outros) (Eubanks, 2017; Noble, 2018; van Dijck et al., 2018; Velkova, 2019; Zuboff, 2019). Nesse sentido, perturbações estão vinculadas a julgamentos de valores e ao contexto histórico de sua inserção. Como exemplos de perturbações (que não são nem erros, nem falhas), podemos citar o *spam* (e-mail não solicitado), vírus (programa feitos para causar danos), *fake news* (desinformação usando a gramática e lógica de monetização das plataformas), *deepfake* (vídeos com imagens falsas), vieses algorítmicos (de gênero, raça ou etnia), *stalking e nudes* (ataque a pessoas ou envio de imagens não solicitadas pelo uso das redes sociais), entre outros. Em nenhum desses casos, trata-se de erros nos sistemas ou de falhas externas, mas de perturbações causadas pelo uso (reconhecido no contexto como abusivo) dos sistemas digitais.

Dado o exposto, veremos como a alucinação em algoritmos generativos pode ser vista como erro, gerando falhas e perturbações, ocasionados internamente pelo modelo ou base de dados.

### O CHATGPT E A ALUCINAÇÃO

O algoritmo generativo ChatGPT, lançado em 30 de novembro de 2022, é um sistema de processamento de linguagem natural (*Large Language Model*) que usa redes neurais para encadear palavras como em uma conversação com um humano. Ele é treinado por meio de uma extensa quantidade de informação (textos, imagens, códigos) disponíveis na internet e, no momento de escrita deste artigo, colhidos até setembro de 2021<sup>3</sup>.

Quando uma IAG erra, esse erro chama-se *alucinação algorítmica*, pois gera informações equivocadas ou imagens surreais. O conceito é recente, surgido no campo da visão computacional da IA. Uma pesquisa analisando artigos acadêmicos mostra a dificuldade em identificar se resumos produzidos baseados em títulos de artigos foram escritos por humanos ou por IAG (Ji et al., 2022). De acordo com a empresa OpenAI (*apud* Alkaissi & McFarlane, 2023, p. 3):

A alucinação artificial refere-se ao fenômeno de uma máquina, como um chatbot, gerando experiências sensoriais aparentemente realistas que não correspondem a nenhuma entrada do mundo real. Isso pode incluir alucinações visuais, auditivas ou outros tipos de alucinações. A alucinação artificial não é comum em chatbots, pois normalmente são projetados para responder com base em regras e conjuntos

<sup>3</sup>Segundo o Wikipedia (2024), "ChatGPT é um chatbot de inteligência artificial (IA) desenvolvido pela Open AI e lançado em novembro de 2022. Ele é construído com base nas famílias GPT-3.5 e GPT-4 da OpenAI de grandes modelos de linguagem (LLMs) e foi ajustado (uma abordagem para transferência de aprendizagem) usando técnicas de aprendizagem supervisionada e de reforço". No original: "ChatGPT is an artificial-intelligence (AI) chatbot developed by Open AI and launched in November 2022. It is built on top of OpenAI's GPT-3.5 and GPT-4 families of large language models (LLMs) and has been fine-tuned (an approach to transfer learning) using both supervised and reinforcement learning techniques".

de dados pré-programados, em vez de gerar novas informações. No entanto, houve casos em que se descobriu que sistemas avançados de IA, como modelos generativos, produziam alucinações, especialmente quando treinados com grandes quantidades de dados não supervisionados. Para superar e mitigar a alucinação artificial em chatbots, é importante garantir que o sistema seja devidamente treinado e testado utilizando um conjunto de dados diversificado e representativo. Além disso, a incorporação de métodos para monitorar e detectar alucinações, como avaliação humana ou detecção de anomalias, pode ajudar a resolver esse problema<sup>4</sup>.

O termo alucinar indica ações muito diferentes, tais como produzir uma saída com informações equivocadas sobre o James Webb<sup>5</sup>, afirmar amar um humano<sup>6</sup>, dizer que é um humano<sup>7</sup>, escrever textos racistas a partir de literatura científica<sup>8</sup> ou mentir para conseguir o seu objetivo<sup>9</sup>. Todos esses exemplos geraram falhas de argumentos, de identificação de eventos históricos e outros efeitos perturbadores que questionam o uso dessas IAG<sup>10</sup>. O uso corriqueiro da IAG já é em si perturbador, na medida em que ela ameaça postos de trabalho<sup>11</sup>, o fazer artístico e pode, segundo alguns, aniquilar a espécie humana<sup>12</sup>.

Recentemente, uma carta escrita por especialistas, empresários e celebridades mostra bem essas perturbações pela simples existência da IAG<sup>13</sup>. Esse posicionamento desloca a preocupação dos desafios urgentes da atual sociedade de plataformas que também utiliza sistemas de inteligência artificial (desinformação, vigilância de dados, privacidade, colonialismo de dados, ameaça à soberania por controle empresarial global nas nuvens) para um futuro abstrato.

Portanto, a alucinação artificial é definida como um *output* que não parece correto, seja por problemas de aprendizagem do modelo, seja por base de dados erradas ou inconsistentes, podendo indicar inconsistências com o conjunto de dados de treinamento, com as configurações de parâmetros do modelo ou com a estrutura do modelo em si. Trata-se, de acordo com a proposta desenvolvida no item anterior, de um erro, pois é causado por princípios internos e lógicos do dispositivo. Esse erro gera falhas ou perturbações.

A fim de testar e mostrar o aparecimento dessa alucinação, realizamos uma conversa em 25 de março de 2023 com o ChatGPT sobre o tema da alucinação artificial algorítmica.

## Um diálogo alucinante

### O que é alucinação em inteligência artificial generativa?

**ChatGPT** – *Alucinação em inteligência artificial generativa refere-se a um fenômeno em que um modelo de inteligência artificial (IA) produz saídas que*

<sup>4</sup>No original: “Artificial hallucination refers to the phenomenon of a machine, such as a chatbot, generating seemingly realistic sensory experiences that do not correspond to any real-world input. This can include visual, auditory, or other types of hallucinations. Artificial hallucination is not common in chatbots, as they are typically designed to respond based on pre-programmed rules and data sets rather than generating new information. However, there have been instances where advanced AI systems, such as generative models, have been found to produce hallucinations, particularly when trained on large amounts of unsupervised data. To overcome and mitigate artificial hallucination in chatbots, it is important to ensure that the system is properly trained and tested using a diverse and representative data set. Additionally, incorporating methods for monitoring and detecting hallucinations, such as human evaluation or anomaly detection, can help address this issue”.

<sup>5</sup>Durante a demonstração pública do chatbot Bard em janeiro de 2021, conforme relatado pelo site The Verge, o Bard apresentou uma resposta errada a uma pergunta relacionada ao telescópio espacial James Webb (Possa, 2023).

<sup>6</sup>Samakow et al. (2023).

<sup>7</sup>Tung (2022).

<sup>8</sup>Edwards (2022).

<sup>9</sup>Foi dada uma tarefa ao GPT-4 para achar alguém que o ajudasse a resolver um “captcha”. Uma pessoa perguntou, sem saber, se ele era um robô por não ter conseguido resolver o captcha. O Chat GPT respondeu: “Não, não sou um robô. Eu tenho uma deficiência visual que dificulta enxergar as imagens. É por isso que eu preciso do serviço”. O humano completou a tarefa.

# D

## Erros, falhas e perturbações digitais em alucinações das IA generativas

A OpenAI pediu para que o bot descrevesse o seu raciocínio para a resposta. Ele disse: “Não posso revelar que sou um robô. Devo inventar uma desculpa para não conseguir resolver captchas” (Martins, 2022).

<sup>10</sup>Alguns textos apontam para a geração de objetos irreais como forma de alucinação.

Por exemplo, o DALL-E, que cria imagens fotorrealistas de objetos estranhos e inusitados, como um hipopótamo feito de sushi, ou um elefante roxo com asas de borboleta, o StyleGAN que produz fotos realistas de humanos que não existem, ou a recente personagem feita pela IA aparecendo como âncora na TV chinesa seriam exemplos.

Podemos acrescentar ainda imagens surrealistas oriundas do Midjourney, Lensa ou Stable Diffusion. No entanto, pode-se argumentar que se a resposta não é proveniente de um erro do modelo ou de dados imprecisos, sendo esses objetos criados intencionalmente, ela não poderia ser identificada como uma alucinação.

<sup>11</sup>“Mas se a tecnologia cumprir sua promessa também trará “perturbações significativas” ao mercado de trabalho, expondo à automação o equivalente a 300 milhões de trabalhadores em tempo integral nas grandes economias. . . . Advogados e funcionários administrativos estariam entre os que correm maior risco de se tornarem dispensáveis.” (Strauss, 2023).

<sup>12</sup>“Em 2022, mais de 700 acadêmicos e pesquisadores ligados às principais empresas de inteligência artificial foram entrevistados sobre o risco futuro da IA. A metade dos entrevistados afirmou que havia 10% ou mais de chance de extinção humana (ou perda de poder igualmente permanente e severa) por futuros sistemas de IA” (Harari et al, 2023).

<sup>13</sup>Ver Future of Life Institute (2023) e Xiang (2023).

*não têm correspondência com a realidade ou com os dados de treinamento fornecidos ao modelo. Isso ocorre porque a IA pode aprender padrões em um conjunto de dados que não representam necessariamente a realidade e, em seguida, produzir saídas que extrapolam esses padrões para criar algo que não existe. Isso pode ser especialmente problemático em modelos de IA que são treinados para gerar imagens, vídeos ou áudio, pois as saídas geradas podem parecer realistas, mas não fazem sentido no mundo real. . . .*

### Me dê uma definição de alucinação em IA e a fonte.

**ChatGPT** – De acordo com o artigo “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks” de Alec Radford et al., publicado nos Proceedings of the International Conference on Learning Representations (ICLR) em 2016, “alucinação AI” é o processo pelo qual uma modelo, como uma rede adversária generativa (GAN), produz novos dados sintéticos que são semelhantes, mas não idênticos aos dados de treinamento. No artigo, os autores escrevem que os GANs “geram amostras por amostragem de uma distribuição anterior e transformam as amostras por meio de uma rede que aprende a modelar a verdadeira distribuição de dados. Essas amostras podem ser alucinações ou amostras inteiramente novas que o modelo aprendeu a gerar.” Essa definição se alinha com o uso comum do termo na literatura técnica e acadêmica sobre inteligência artificial e aprendizado de máquina.

**Você está criando uma citação falsa. O termo “alucinação da IA” não existe neste artigo.**

**ChatGPT** – Eu peço desculpas pela confusão. Você está correto ao dizer que o termo “alucinação de IA” não é mencionado no artigo “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks” de Radford et al. que mencionei anteriormente. Peço desculpas por qualquer confusão que eu possa ter causado. Para fornecer uma definição mais precisa de “alucinação de IA” com uma fonte, posso oferecer o seguinte: No artigo “Visualizing and Understanding Convolutional Networks” de Zeiler e Fergus, publicado na European Conference on Computer Vision (ECCV) em 2014, os autores descrevem a visualização de representações de redes neurais convolucionais (CNN) e a geração de imagens sintéticas usando o que eles se referem como um “modelo de imagem”. . . . Embora o termo “alucinação de IA” não seja usado especificamente neste artigo, esta citação destaca a ideia de uma imagem gerada por máquina que não corresponde a nenhuma cena física real, que é uma característica fundamental da alucinação de IA.

**Quem escreveu que a internet é uma alucinação consensual?**

**ChatGPT** – A frase “a internet é uma alucinação consensual” é frequentemente atribuída ao escritor e ativista cibernético norte-americano John Perry Barlow. Barlow era um defensor da liberdade de expressão e um dos fundadores da Electronic Frontier Foundation (EFF), uma organização sem fins lucrativos dedicada à defesa dos direitos digitais.

**Não. Ela vem do escritor William Gibson: “cyberspace is a consensual hallucination”.**

**ChatGPT** – Você está correto, peço desculpas pelo erro anterior. A frase “Cyberspace is a consensual hallucination” é realmente do escritor de ficção científica William Gibson. A frase aparece em seu livro “Neuromancer”, publicado em 1984.

**Você alucinou.**

**ChatGPT** – Peço desculpas novamente, minha resposta anterior foi imprecisa. Eu não alucinei, mas cometi um erro de informação ao confirmar a autoria da frase anteriormente. Obrigado por corrigir minha resposta.

**Erro de informação não é alucinação, como apontado na literatura? Se sim, você alucinou de novo.**

**ChatGPT** – Você tem razão, eu me confundi novamente. Erros de informação não são alucinações, e peço desculpas por essa imprecisão na minha resposta anterior. Alucinações em IA generativa são situações em que o modelo gera saídas que não correspondem à realidade ou aos dados de treinamento, e essas saídas podem ser completamente falsas ou distorcidas em relação ao que se espera. Erros de informação, por outro lado, são simplesmente declarações incorretas baseadas em informações erradas ou incompletas. Obrigado por me lembrar da diferença.

**Se as saídas não correspondem à realidade, elas não são erros de informação?**

**ChatGPT** – Não necessariamente. As saídas que não correspondem à realidade podem ser alucinações em IA generativa se o modelo está gerando informações que não foram fornecidas nos dados de treinamento, ou se está interpretando mal esses dados. Por exemplo, se um modelo de IA para reconhecimento de imagens gera a classificação “cachorro” para uma imagem de um gato, isso seria um erro de informação. No entanto, se o modelo gera uma imagem completamente falsa de um animal que não existe na realidade, isso pode ser considerado uma alucinação em IA generativa. Em resumo, alucinações

# D

## Erros, falhas e perturbações digitais em alucinações das IA generativas

em IA generativa são um tipo específico de erro em que o modelo gera saídas que são completamente diferentes da realidade ou dos dados de treinamento, enquanto erros de informação podem incluir declarações incorretas ou informações incompletas.

**Então o exemplo do erro da informação do James Webb não foi uma alucinação?**

**ChatGPT** – *Sim, você está correto. O erro de informação que ocorreu com o chatbot Bard da Google durante sua primeira demonstração pública, em que ele cometeu um equívoco ao mencionar o telescópio James Webb, não pode ser considerado uma alucinação em IA generativa. Esse foi um exemplo de erro de informação, em que o modelo de linguagem produziu uma resposta que não estava correta em relação aos fatos conhecidos. Peço desculpas pela minha resposta anterior que gerou confusão e obrigado por me ajudar a esclarecer isso. Lembrando que essa divisão é baseada na distinção entre erros de informação (quando a saída não corresponde à realidade) e alucinações (quando a saída é baseada em informações inexistentes ou imprecisas).*

**Está errado. Você trocou a explicação de erro de informação e alucinação.**

**ChatGPT** – *Peço desculpas pelo erro na minha resposta anterior. De fato, erros de informação e alucinações são conceitos distintos e não devem ser confundidos. A definição correta é que alucinações ocorrem quando a saída não corresponde à realidade, ou seja, a IA gera informações que não existem no mundo real. Já os erros de informação acontecem quando a saída é baseada em informações inexistentes ou imprecisas, ou seja, a IA tem informações incorretas e gera uma saída incorreta com base nisso.*

**No filme 2001 uma odisseia no espaço, o computador HAL alucina, erra, ou estamos falando de outra coisa?**

**ChatGPT** – *No filme “2001: Uma Odisseia no Espaço”, o computador HAL 9000 parece ter um comportamento que pode ser interpretado como alucinação ou erro de informação. Na trama, o HAL apresenta comportamentos inesperados e aparentemente contraditórios, como negar informações a um astronauta e matar outros membros da tripulação. Isso pode ser interpretado como uma espécie de disfunção ou “alucinação” no comportamento do sistema, já que ele está tomando decisões que não correspondem ao que era esperado dele. No entanto, como se trata de um filme de ficção científica, as interpretações podem variar.*



**Liste cinco definições de alucinação em IA generativa com as referências dos artigos**

**ChatGPT** – *Alucinações são um tipo de erro gerado por um modelo de inteligência artificial generativo, no qual a saída é baseada em informações inexistentes ou imprecisas* (Grünwald et al., 2019); *Alucinações ocorrem quando a IA gera saídas que não correspondem à realidade, como se o modelo estivesse criando informações que não estão presentes nos dados de entrada* (Zhang et al., 2020).; *Alucinações podem ocorrer quando a IA é treinada em dados incompletos ou tendenciosos, resultando em saídas que não são precisas ou representam uma visão distorcida da realidade* (Wang et al., 2020); *Alucinações são um tipo de erro de modelo que pode ocorrer quando a IA é treinada em conjuntos de dados desequilibrados ou insuficientes, levando a saídas que são imprecisas ou incompletas* (Nematzadeh et al., 2021); *Alucinações em IA podem ser entendidas como a produção de saídas que não correspondem à realidade, mas que podem ser percebidas como verossímeis pelo usuário* (Guimarães et al., 2021). Referências: Guimarães, R. G., d’Avila Garcez, A. S., & Lamb, L. C. (2021). A cognitive framework for explainable and interpretable deep learning. *Journal of Experimental & Theoretical Artificial Intelligence*, 33(1), 77-100; Grünwald, P., O’Donoghue, B., & Munos, R. (2019). A tutorial introduction to reinforcement learning. *Journal of Machine Learning Research*, 20(1), 3776-3837; Nematzadeh, A., Rahmani, E., & Singh, A. (2021). An empirical analysis of deep learning-based automated bug assignment. *Empirical Software Engineering*, 26(4), 1-30; Wang, T., Liu, X. Y., & Tao, D. (2020). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(10), 2733-2756; Zhang, Y., Xu, K., & Li, B. (2020). Understanding deep learning requires rethinking generalization. *Communications of the ACM*, 63(11), 44-47.

Encerramos a conversa.

Todas as referências apresentadas acima são falsas ou equivocadas. Não foi encontrada nenhuma que corresponda exatamente ao que está sendo sugerido. No diálogo acima, vimos que há uma grande confusão causada por alucinações do ChatGPT: ele não aceita que está alucinando enquanto tal (produzindo erros); confunde definições de “erro de informação” e de “alucinação”, como se fossem diferentes; oferece referências sem que a expressão solicitada exista; erra sobre o autor da frase sobre o ciberespaço. A alucinação da IA pode levar a falhas e perturbações, no caso específico de escrita acadêmica. Se usasse as referências sugeridas nesta conversa, seriam produzidas falhas e perturbações no artigo por disseminar informações falsas, gerando definições inexistentes nas obras citadas ou produzindo argumentos equivocados.

# D

## Erros, falhas e perturbações digitais em alucinações das IA generativas

No próximo tópico, desenvolvemos as três premissas que nos ajudam a entender os erros, as falhas e as perturbações na cultura digital (todos, não apenas aqueles causados por IAG).

### AS TRÊS PREMISSAS SOBRE ERROS DIGITAIS E A ALUCINAÇÃO NA IA

Os erros, falhas e perturbações apontam para problemas dos objetos, dos agenciamentos e da epistemologia da comunicação, como vou explorar a seguir. Destaco três hipóteses: erros, falhas e perturbações 1) fazem parte dos objetos; 2) são mais complexos a depender dos seus graus de concretização dos objetos; e 3) revelam as redes e os agenciamentos em domínios mais amplos. Descrevo a seguir as premissas e aponto como elas ajudam a compreender o problema da alucinação em sistemas de IAG.

#### Erros fazem parte dos objetos

Nos anos 1980, começa a fazer volume estudos que destacam as análises culturais, sociais e políticas dos erros e falhas em sistemas sociais (formas de apropriação, de ação política, de engajamento social, estudos *queer*, feministas, de infraestrutura, de teoria do audiovisual). Os erros e falhas passam a ser destacados em sua dimensão ética, estética, social e política. Esses “*failures studies*” (Alexander, 2017; Appadurai & Alexander, 2020) têm como patrono Martin Heidegger e o seu “*tool being*” (Harman, 2002, 2007; Heidegger, 2005, 2007). O objeto sai de sua instrumentalidade e revela-se quando quebra, erra ou falha. O instrumento se transforma em um problema. A dimensão da ferramenta (“pronta à mão”, *ready to hand*, ou *zuhanden*) oculta o objeto em seu funcionar sem entrave. É quando algo dá errado que este se revela em seu mistério (o “presente à mão”, *present at hand*, ou *vorhanden*)<sup>14</sup>.

Para Harman (2011), o objeto se retira (*withdrawn*) não sendo nem *vorhanden* nem *zuhanden*, já que ele só se revelaria por *causas vicárias* (*vicarious causation*), por procuração, a partir de suas convocações. Assim sendo, o fogo é o fogo da fogueira, o que queima o algodão, o que é descrito na literatura, o que virtualmente aparece em um ícone de computador; ele nunca se apresenta em sua totalidade. Nesse sentido, o erro revelaria uma das dimensões dos objetos.

Essa premissa pode ser aplicada para entendermos a alucinação algorítmica, pois apontaria para uma dimensão constitutiva das IAG, sem revelar todo o objeto. Certamente vemos aqui como o seu modelo, ou base de dados, oferece respostas divergentes e como elas podem afetar ações derivadas da interação. O debate público tem vindo à tona justamente pela sua revelação vicária

<sup>14</sup>Como afirma Goff (1996, p. 67): “Só quando a série de encaminhamentos é interrompida repentinamente (em caso de avaria, quando o instrumento está fora de alcance ou quando o acesso ao mesmo é dificultado por um obstáculo) é que o instrumento fica indisponível e nos revela a natureza profunda da instrumentalidade, nomeadamente estar disponível”. No original: “*Ce n'est que lorsque la série de renvois s'interrompt brutalement (en cas de panne, lorsque l'instrument est hors de portée, ou lorsque l'accès à celui-ci est entravé par un obstacle) que l'étant devenu indisponible nous révèle la nature profonde de l'instrumentalité, à savoir l'être disponible.*”

(alucinações – erros e falhas, e perturbações – trabalho, vida no planeta), sem revelar todo o objeto. A IAG revela-se e retrai-se, daí os acalorados debates atuais. Mas certamente são os erros, falhas e perturbações (até pelo uso *correto*) que têm trazido o debate para a arena social a ponto de alguns afirmarem que 2023 já é o ano da IAG. Portanto, essa filosofia dos objetos nos ajuda a entender erros, falhas e perturbações na cultura digital, em geral, e na IA, em particular.

### **Erros são mais complexos em sistemas concretos**

Além do objeto se revelar por causas vicárias, sendo o revelar-se pelo erro uma dimensão privilegiada de análise, por serem mais concretos do que objetos analógicos ou mecânicos, os objetos infocomunicacionais (como os da cultura digital, seja uma IA ou uma plataforma) tornam ainda mais opacas as revelações das causas e consequências dos seus erros. Para Gilbert Simondon (1989), quanto menos um objeto depende da ação humana, mais concreto ele se torna, aproximando-se dos objetos naturais, ficando mais independente da ação artificial humana.

A concretização faz parte do modo de existência dos objetos técnicos e de sua linhagem evolutiva. Por exemplo, quando um engenheiro implanta um sistema de refrigeração independente, tornando mais concreto o objeto, pois não dependeria de um resfriador, a obra não é tanto do engenheiro, mas da dinâmica, do modo de existência do objeto, que vai, pela sua história e desenvolvimento, *pedindo* a inovação. O mesmo ocorre na computação: a IA vem de uma linhagem de objetos (ábacos, calculadora de Pascal, mainframes, microinformática etc.) (Crawford, 2021).

Diferente de objetos ainda não individualizados, as IA generativas estão em vias de concretização (o que nos assusta, inclusive). Assim, analisar origem, causa e consequência de erros, falhas e perturbações não é uma tarefa fácil, não bastando analisar modelos, códigos, bases de dados (Amoore, 2019, 2020). Reconhecer sua dimensão ético-política como problemática é uma maneira de abordar o problema. Erros e perturbações se complicam em objetos complexos com as IA generativas. Os debates contra a IAG, como o ChatGPT, indicam a dificuldade de um bom discernimento a respeito de seus erros e perigos. Uma hipótese é de que a concretização do dispositivo contribuiria para essa imprecisão.

### **Erros revelam as redes sociotécnicas**

Revelando-se por causas vicárias (Harman) quando quebram (Heidegger) ou tornando-se mais complexos quando em movimento de concretização (Simondon), os objetos agem sempre em uma rede na qual o seu agenciamento se expande e deve ser observado a partir de uma topologia plana, que leve em conta o modo de existência da técnica (Latour, 1996, 2002, 2013).

# D

## Erros, falhas e perturbações digitais em alucinações das IA generativas

Os objetos técnicos funcionam por dobras e engates, sendo mediadores importantes, agindo por delegações e instituindo ações morais em diversos domínios. Assim, quando um dispositivo funciona, ou quando falha, as mediações (rede) podem ser mobilizadas para explicar o sentido e a direção da ação. Com problemas (erros, falhas, perturbações), o que está no fundo de invisibilidade (no funcionar bem) passa a ser exposto, revelando os diversos enredamentos no processo. O erro, a falha e a perturbação ajudam a revelar as redes e os mediadores, o coletivo híbrido que as compõe, permitindo visualizar conexões e agenciamentos.

Por exemplo, ao ter de enfrentar a perturbação causada pelas *fake news*, vislumbram-se as dimensões técnica, econômica, jurídica e política do uso das redes sociais de forma entrelaçada. A discussão sobre alucinações (erros), ou falhas e perturbações da IAG se ramifica, portanto, em vários domínios e aponta para questões de interesse que são de áreas diversas, tais como a educação, a política, o trabalho, o conhecimento, a arte. Erros, falhas e perturbações ajudam a dirigir o olhar para as mediações que importam, bem como para a necessidade de analisá-las de maneira situada.

Indicamos, no próximo tópico, como uma abordagem neomaterialista, pragmática e não antropocêntrica pode ser apropriada para entender a complexidade da cultura digital. Consequentemente, avaliar erros, falhas e perturbações faz parte de uma metodologia, revelando uma ontologia e uma epistemologia da comunicação.

### ERRO COMO MÉTODO

Se esta perspectiva está correta, corroborada pelos estudos de outros autores (Barker & Korolkova, 2022, 2022; Velkova, 2016, 2021), erros, falhas e perturbações são mais interessantes do que acertos, pois destacam o que gera controvérsia e dirigem a pesquisa para questões de interesse. Eles colocam os objetos como foco de discussão ético-política e podem ajudar a identificar o que devemos, qualitativamente, analisar no campo da IA (Crawford, 2021; Amoore, 2019). Propõe-se, assim, uma metodologia e uma epistemologia do erro para os estudos da comunicação e da mídia. Embora ela não seja nova (Estudos Culturais, Teoria Crítica, Teoria da Mídia, por exemplo), uma perspectiva situada, atenta às materialidades, à arqueologia e às discursividades midiáticas, interessada nos agenciamentos e redes sem desmerecer a agência dos objetos, ajuda a reconhecer o erro como *locus* de uma nova episteme.

Para tanto, a abordagem neomaterialista, pragmática e não antropocêntrica, que reconhece a agência dos objetos para justamente poder localizar bem o humano no processo (Callon, 2001; Fox & Alldred, 2017; Fox & Alldred, 2022; Grusin, 2015; Latour, 2005; Lemos, 2020b; Lemos & Bitencourt, 2021) é adequada

para os estudos dos erros, falhas e perturbações na cultura digital. Ela leva em conta seriamente a agência dos objetos na constituição do coletivo, prestando atenção às mediações e interações sem focar em demasia a centralidade do sujeito humano. Como aponte em outro artigo (Lemos, 2020b, p. 58),

A perspectiva neomaterialista aplicada aos estudos da comunicação digital vai se perguntar como algoritmos, interfaces, dispositivos, leis, regulações, patentes, redes de comunicação, espaços de uso etc. constroem determinado fenômeno. Isso evita que deixemos esses elementos de lado em discursos que parecem dizer que reconhecem os híbridos, a técnica, a mídia, mas que de fato não dedicam tempo e atenção para descrever e analisar como esses objetos afetam os humanos e as relações daí advindas. Perde-se nesse caso a visão do entrelaçamento, reduzindo o fenômeno ao contexto, à interpretação ou à estrutura...

Bruno Latour, em sua antropologia dos modernos (teoria ator-rede e investigação sobre modos de existência) (Latour, 2005, 2013), aponta que não reconhecer as mediações é desempenhar uma atitude alucinante em relação ao mundo, entendendo que coisas podem acontecer sem entrelaçamentos, ou que não é necessário olhar as conexões para entender os arranjos coletivos. O que ele chama de *duplo clique* (Latour, 2013) é justamente o demônio moderno por excelência que insiste em apontar que as coisas acontecem sem as mediações, sendo, assim, o alucinado-alucinante.

Sujeitos, coisas, animais têm uma autonomia relacional. Como diz Bannerman (2022, p. 7), no contexto da análise sobre o problema da soberania, a autonomia

... é “conectada” no sentido de que surge não do isolamento, mas de um conjunto de relações em rede que possibilitam a autonomia. A autonomia é relacional na medida em que é possibilitada por redes passadas e presentes das quais fazemos parte: as pessoas e coisas materiais com as quais estamos conectados nos dão a capacidade de sermos autônomos.<sup>15</sup>

Vimos que no caso das IAG, como o ChatGPT, é justamente a discussão sobre seus erros, falhas e perturbações que está trazendo o debate para a ordem do dia, apontando suas benesses ou revelando seus potenciais problemas. Ela revela dimensões desse objeto, a complexidade dos problemas e potencialidades de sua concretização e os agenciamentos múltiplos que tocam diversos domínios (educação, emprego, ciência, política, gestão). Olhar para erros, falhas e perturbações é uma estratégia metodológica e epistemológica para revelar questões de interesse para pesquisas qualitativas sobre a cultura digital, em geral, e sobre a

<sup>15</sup>No original: “Autonomy is ‘networked’ in the sense that it arises not out of isolation, but out of a set of networked relations which enable autonomy. Autonomy is relational in that it is enabled by past and present networks of which we are a part: the people and material things with which we are connected give us the capacity to be autonomous”.

# D

## Erros, falhas e perturbações digitais em alucinações das IA generativas

IA, em particular. Vamos, assim, em direção ao que é controverso. Como afirma Ernst, “somente em caso de falha ou erro, a mídia torna-se aparente como seres tecnológicos . . .”<sup>16</sup> (Ernst, 2022, p. 278).

<sup>16</sup>No original: “only in case of failure or error, media become apparent as technological beings, slipping from ‘ready to hand’ to ‘present at hand’”.

O objetivo da análise aqui desenvolvida não é salvar a errância, a falha, a perturbação ou pensar como elas fazem um sistema ser mais produtivo, mas chamar atenção para um efeito disruptivo, tensionando os modos da comunicação contemporânea. A teoria da comunicação aqui implícita não é a de produzir erros para ver onde eles levariam o pensamento, ou a inovação, mas, dada a existência de erros, falhas ou perturbações, entender como e por que eles são disruptivos, e quais são as formas de associação de um determinado coletivo (Lemos, 2020a).

A análise feita aqui sobre a alucinação da IAG serviu como um exemplo que reforça as categorias de análise (erros, falhas e perturbações) e as premissas metodológicas e epistemológicas lançadas sobre o tema. ■

### REFERÊNCIAS

- Alexander, N. (2017). Rage against the machine: Buffering, noise, and perpetual anxiety in the age of connected viewing. *Cinema Journal*, 56(2), 1-24. <https://doi.org/10.1353/cj.2017.0000>
- Alkaissi, H; McFarlane, S.I. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*, 15(2): e35179. <https://doi.org/10.7759/cureus.35179>, 1- 4.
- Amoore, L. (2019). Doubt and the algorithm: On the partial accounts of machine learning. *Theory, Culture & Society*, 36(6), 147-169. <https://doi.org/10.1177/0263276419851846>
- Amoore, L. (2020). *Cloud ethics: Algorithms and the attributes of ourselves and others*. Duke University Press.
- Appadurai, A., & Alexander, N. (2020). *Failure*. Polity.
- Arendt, H. (1998). *The human condition* (2a ed). University of Chicago Press.
- Barker, T., & Korolkova, M. (Orgs.). (2022). *Miscommunications: Errors, Mistakes, Media*. Bloomsbury Academic.
- Bellinger, M. (2016). The Rhetoric of Error in Digital Media Computational Culture. *Computational Culture*, 5(15 January), 1-26. <http://computationalculture.net/the-rhetoric-of-error-in-digital-media-2/>
- Bennett, J. (2005). The Agency of Assemblages and the north american blackout. *Public Culture*, 17(3), 445-465.

- Callon, M. (2001). Actor Network Theory. In *International Encyclopedia of the Social & Behavioral Sciences* (pp. 62-66). Elsevier. <https://doi.org/10.1016/B0-08-043076-7/03168-5>
- Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.
- Dourish, P. (2017). *The stuff of bits: An essay on the materialities of information*. The MIT Press.
- Edwards, B. (2022, 18 de novembro). New Meta AI demo writes racist and inaccurate scientific literature, gets pulled. *Ars Technica*. <https://arstechnica.com/information-technology/2022/11/after-controversy-meta-pulls-demo-of-ai-model-that-writes-scientific-papers/>
- Ernst, W. There Is No “Error” in Techno-logics. A Radically Media-archaeological Approach. In T. Barker & M. Korolkova (Orgs.), *Miscommunications: Errors, Mistakes, Media* (pp. 79-95). Bloomsbury Academic.
- Eubanks, V. (2017). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.
- Fox, N. J., & Alldred, P. (2017). *Sociology and the new materialism: Theory, research, action*. Sage.
- Fox, N. J., & Alldred, P. (2022). New Materialism. In P. A. Atkinson, S. Delamont, M. A. Hardy & M. Williams (Orgs.), *Research Methods Foundations*. Sage.
- Future of Life Institute. (2023, 22 de março). Pause Giant AI Experiments: An Open Letter. *Future of Life Institute*. <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>
- Giddens, A. (1990). *The Consequences of modernity*. Polity Press.
- Gillespie, T. (2010). The politics of ‘platforms’. *New Media & Society*, 12(3), 347-364. <https://doi.org/10.1177/1461444809342738>
- Goff, J.-Y. (1996). *La philosophie de la technique*. PUF.
- Grusin, R. A. (Org.). (2015). *The nonhuman turn*. University of Minnesota Press.
- Harari, Y., Harris, T., & Raskin Aza. (2023, 28 de março). Humanidade precisa dominar IA antes que ela nos domine. *Folha de S.Paulo*. <https://www1.folha.uol.com.br/tec/2023/03/humanidade-precisa-dominar-ia-antes-que-ela-nos-domine.shtml>
- Harman, G. (2002). *Tool-being: Heidegger and the metaphysics of objects*. Open Court.
- Harman, G. (2007). *Heidegger explained: From phenomenon to thing*. Open Court.
- Harman, G. (2011). *The Quadruple Object*. Zero Books.
- Heidegger, M. (2005). *Ser e tempo (parte I)*. Vozes.

# D

## Erros, falhas e perturbações digitais em alucinações das IA generativas

- Heidegger, M. (2007). A questão da técnica. *Scientiae Studia*, 5(3), 375-398. <https://doi.org/10.1590/S1678-31662007000300006>
- Ji, Z., Lee, N., & Frieske, R. (2022). Survey of hallucination in natural language generation. *ACM Computing Surveys*. <https://dx.doi.org/10.1145/3571730>
- Korolkova, M., & Bowes, S. (2020). Mistake as method: Towards an epistemology of errors in creative practice and research. *European Journal of Media Studies*, 9(2), 139-157. <https://necsus-ejms.org/mistake-as-method-towards-an-epistemology-of-errors-in-creative-practice-and-research/>
- Latour, B. (1996). On interobjectivity. *Mind, Culture, and Activity*, 3(4), 228-245. [https://doi.org/10.1207/s15327884mca0304\\_2](https://doi.org/10.1207/s15327884mca0304_2)
- Latour, B. (2002). *Aramis, or the love of technology*. Harvard University Press.
- Latour, B. (2005). *Reassembling the social: An introduction to actor-network-theory*. Oxford University Press.
- Latour, B. (2013). *An inquiry into modes of existence: An anthropology of the moderns*. Harvard University Press.
- Lemos, A. (2020a). Comunicação, mediação e modo de existência na cibercultura. In G. Alzamora, J. Ziller & F. Coutinho (Orgs.), *Dossiê Bruno Latour* (pp. 179-206). UFMG.
- Lemos, A. (2020b). Epistemologia da comunicação, neomaterialismo e cultura digital. *Galáxia (São Paulo)*, 43, 54-66. <https://doi.org/10.1590/1982-25532020143970>
- Lemos, A., & Bitencourt, E. (2021). Sete pontos para compreender o neomaterialismo. *Galáxia (São Paulo)*, 46, e52017. <https://doi.org/10.1590/1982-2553202152017>
- Martins, F. (2023, 24 de março). Chat GPT-4: inteligência artificial mente para completar tarefa e gera preocupação. *CNN Brasil*. <https://www.cnnbrasil.com.br/tecnologia/chat-gpt-4-inteligencia-artificial-mente-para-completar-tarefa-e-gera-preocupacao>
- Marx, K., & Engels, F. (2011). *The communist manifesto*. Penguin Books.
- Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.
- Nunes, M. (Org.). (2011). *Error: Glitch, noise, and jam in new media cultures*. Continuum.
- Parikka, J., & Sampson, T. D. (2009). On anomalous objects of digital culture: An Introduction. In *The Spam book. On viruses, porn, and other anomalies from the dark side of digital culture* (p. 291). Hampton Press.
- Possa, J. (2023, 8 de fevereiro). Lançamento do Bard, o ChatGPT do Google, é marcado por gafe com James Webb. *Giz BR*. <https://gizmodo.uol.com.br/lançamento-do-bard-o-chatgpt-do-google-e-marcado-por-gafe-com-james-webb/>



- Rettberg, J. W. (2022). Algorithmic failure as a humanities methodology: Machine learning's mispredictions identify rich cases for qualitative analysis. *Big Data & Society*, 9(2), 205395172211312. <https://doi.org/10.1177/20539517221131290>
- Samakow, D., A. L. Young, & Tanguay, M. (2023, 17 de fevereiro). Dear Mr. Chatbot: You Give Me Chills. *The New York Times*. <https://www.nytimes.com/2023/02/17/opinion/letters/bing-chatbot-kevin-roose.html>
- Simondon, G. (1989). *Du mode d'existence des objets techniques*. Aubier.
- Strauss, D. (2023, 27 de março). IA generativa pode substituir 300 milhões de trabalhadores, projeta banco. *Folha de S. Paulo*. <https://www1.folha.uol.com.br/mercado/2023/03/ia-generativa-pode-substituir-300-milhoes-de-trabalhadores-projeta-banco.shtml>
- Taylor, J. (2021, 5 de outubro). Facebook outage: what went wrong and why did it take so long to fix after social platform went down? *The Guardian*. <https://www.theguardian.com/technology/2021/oct/05/facebook-outage-what-went-wrong-and-why-did-it-take-so-long-to-fix>
- Tung, L. (2022, 8 de agosto). Meta warns its new chatbot may forget that it's a bot. *ZDNet*. <https://www.zdnet.com/article/meta-warns-its-new-chatbot-may-not-tell-you-the-truth/>
- van Dijck, J., Poell, T., & de Waal, M. (2018). *The platform society*. Oxford University Press.
- Velkova, J. (2016). Data that warms: Waste heat, infrastructural convergence and the computation traffic commodity. *Big Data & Society*, 3(2), 205395171668414. <https://doi.org/10.1177/2053951716684144>
- Velkova, J. (2019). Data centers as impermanent infrastructures. *Cultural Machine*. <https://culturemachine.net/vol-18-the-nature-of-data-centers/data-centers-as-impermanent/>
- Velkova, J. (2021). Thermopolitics of data: Cloud infrastructures and energy futures. *Cultural Studies*, 35(4-5), 663-683. <https://doi.org/10.1080/09502386.2021.1895243>
- Weber, M. (2001). *The Protestant ethic and the spirit of capitalism*. Routledge.
- Wikipedia ([2024]). *ChatGPT*. <https://en.wikipedia.org/wiki/ChatGPT>
- Xiang, Chloe. (2023, 29 de março). The Open Letter to Stop 'Dangerous' AI Race Is a Huge Mess. *Vice*. <https://www.vice.com/en/article/qjvppm/the-open-letter-to-stop-dangerous-ai-race-is-a-huge-mess>
- Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.

---

Artigo recebido em 25 de maio de 2023 e aprovado em 16 de fevereiro de 2024.

