

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

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

■ ANDRÉ LUIZ MARTINS LEMOS^a

Universidade Federal da Bahia, Faculdade de Comunicação, Graduate Program in Contemporary Communication and Culture. Salvador - BA, Brazil

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 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

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,

^aTitular Professor at the Faculty of Communication of the Federal University of Bahia Permanent Professor of the Graduate Program in Contemporary Communication and Culture (Facom/UFBA). CNPQ 1A researcher. Coordinator of Lab404. ORCID <http://orcid.org/0000-0001-9291-6494>. E-mail: almlemos@gmail.com

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

V.18 - Nº 1 jan./abr. 2024 São Paulo - Brasil ANDRÉ LUIZ MARTINS LEMOS p. 75-91

MATRIZES

75



D

Digital errors, failures, and disruptions in generative AI hallucinations

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

DEBATES ABOUT CHATGPT circulate under the sign of error, disturbance, and failure. It has drawn attention for its disruptive nature, whether in “correct” use, threatening jobs and humanity itself, or when it hallucinates, generating false information. By highlighting the benefits of generative artificial intelligence (GenAI), a type of AI that can generate content (data, text, images, sounds) based on learning algorithms, neural networks, and training on a large dataset, the debate is always permeated by its errors, failures, or potential disturbances.

Many studies point to how errors, failures, and disturbances are privileged moments to think about the social and digital culture, in particular (Alexander, 2017; Appadurai & Alexander, 2020; Barker & Korolkova, 2022; Bellinger, 2016; Korolkova & Bowes, 2020; Nunes, 2011; Parikka & Sampson, 2009; Rettberg, 2022). These moments help us understand the life of objects, their agencies, and the forms of realisation and point to a methodological and epistemological dimension of error itself.

Initially, errors, failures, and disturbances in digital culture are defined. Then, ChatGPT and the concept of “algorithmic hallucination” are described. Through a dialogue with this GenAI, we concretely point out how it hallucinates. The third part of the article exposes the premises of research on digital errors and relates them to ChatGPT’s hallucination. The last part addresses the discussion of errors in digital culture as a method to glimpse the agencies at play. It is concluded that errors are not only disruptive or an opportunity to generate innovation but also events that allow understanding of the forms of communication and the actions of digital media.

DEFINITIONS ERRORS, FAILURES AND DISTURBANCES

To understand the problem of error in digital culture, I propose to think in three categories: errors, failures, and disturbances. There is a confusion between the terms. Errors can generate other errors, failures, or disturbances. For example, a misquotation from an AI can lead to argument errors in a scientific article (error). Errors in news (factual) can fuel misinformation (disturbance)! For example, infrastructure failures (electrical or connection) can cause logical system errors and disruptive effects, such as the inability to access banking applications or social networks. Every technological mutation

generates errors, failures, and disturbances, allowing us to understand the place of its production, the historical context, and the social arrangements that shelter it. For a better diagnosis of the disruptive effects of digital culture, I propose to differentiate errors, failures, and disturbances.

Errors are logical problems or principles that deviate from the intended result. The definition is linked to the word's original meaning, wandering, and deviation. Errors are deviant events generated by problems of logic, principle, and internal functioning. As Nunes says, "*error marks a deviation from a predetermined outcome*" (2011, p. 7). Therefore, they are not external phenomena, although they can be generated by them (which we will define as failures), but an integral part of the functioning of cyber-informational systems. An error can produce failures and disturbances, but these are not always generated by it. A logical error can address incorrect information to a user, for example, causing failures and disruptions in the device or the system as a whole. The DNS error (DNS) on Facebook in 2021 generated failures and disturbances as the work via WhatsApp was interrupted in some companies, and people were left without knowing how to act without access to Meta's platforms¹.

¹See (Taylor, 2021).

Failures are problems generated by external events, generating errors and disturbances. They can be caused by logical errors or external phenomena such as infrastructure problems (power outages, cutting off connection cables, access to servers, etc.), accidents or natural disasters, and misuse of systems. Bennett's analysis (2005) of the electrical blackout in the United States 2003 shows how failures generated errors and disturbances, revealing multiple agencies as the cause and consequence of the event.

Disturbances are disruptive events caused mainly by the use in tune with the logic and grammar of the systems or platforms (Dourish, 2017; Gillespie, 2010) and may or may not originate from errors or failures. These disturbances can be seen as anomalies (Parikka & Sampson, 2009) that stress ethical, moral, and political issues according to the historical context. In sociology, disturbances affect a collective ("social system") and trigger changes ("theory of social change") (Arendt, 1998; Giddens, 1990; Marx & Engels, 2011; Weber, 2001). They can have various origins (technological, scientific, political conflicts, economic waves, natural disasters, wars, migrations etc..

Digital (technological) disturbances are caused both by the everyday use of devices (cell phones, computers, and the internet have changed and continue to change society) and by threatening values and social achievements (gender and race bias, environmental issues, work in the *gig economy*; economic regime of control and surveillance, among others) (Eubanks, 2017; Noble, 2018; van Dijck et al., 2018; Velkova, 2019; Zuboff, 2019). In this sense, disturbances

D

Digital errors, failures, and disruptions in generative AI hallucinations

are linked to value judgments and the historical context of their insertion. Examples of disturbances (which are neither errors nor failures) include spam (unsolicited email), viruses (programs designed to cause damage), *fake news* (disinformation using the grammar and logic of platform monetisation), *deepfake* (videos with fake images), algorithmic biases (gender, race, or ethnicity), *stalking*, and *nudes* (attack on people, or sending unsolicited photos through social networks), among others. None of these cases are system errors or external failures but disturbances caused by (recognised in the context as “abusive”) use of digital systems.

Given the above, hallucination in generative algorithms can be seen as an error, generating failures and disturbances as the model or database internally generates it.

CHATGPT AND ALGORITHMIC HALLUCINATION

The generative algorithm ChatGPT, launched on November 30, 2022, is a natural language processing system (*Large Language Model*) that uses neural networks to string words together in conversation like a human. It is trained using extensive information (texts, images, code, etc.) available on the internet (up to September 2021).²

When a GenAI makes a mistake, this error is called *algorithmic hallucination*, generating inaccurate information or surreal images. The concept is recent, emerging in the field of AI computer vision. A study analysing academic articles shows the difficulty of identifying whether abstracts produced based on article titles were written by humans or by GenAI (Ji et al., 2022). According to the company OpenAI (cited in Alkaissi & McFarlane, 2023, p.3):

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 uncommon 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 (p.3).

²Wikipedia ([2024]): 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. <https://en.wikipedia.org/wiki/ChatGPT>

The term hallucination indicates very different actions, such as producing output with inaccurate information about James Webb³, claiming to love a human⁴, saying that it is human⁵, writing racist texts based on scientific literature⁶, or lying to achieve its goal⁷. All these examples have generated failures in arguments, in identifying historical events, and other disturbing effects that question the use of these GenAI⁸. The everyday use of GenAI is already disturbing. It threatens jobs⁹, artistic creation, and can, according to some, annihilate the human species¹⁰.

Recently, a letter written by experts, entrepreneurs, and celebrities demonstrates these disturbances caused by the mere existence of GenAI¹¹. This stance shifts the concern from the urgent challenges of platform societies using artificial intelligence systems (disinformation, data surveillance, privacy, data colonialism, the threat to sovereignty by global corporate control in the clouds...) to an abstract future.

Therefore, artificial hallucination is an output that does not seem correct due to model learning problems or incorrect or inconsistent databases. It may indicate inconsistencies with the training dataset, model parameter settings, or the model structure. According to the proposal developed in the previous section, it is an error, as internal and logical principles of the device cause it. This error generates failures or disturbances. To test and demonstrate the appearance of this hallucination, we conducted a conversation on algorithmic artificial hallucination on March 25, 2023, with ChatGPT.

A hallucinatory dialogue

What is hallucination in generative artificial intelligence?

ChatGPT - *Hallucination in generative artificial intelligence refers to a phenomenon in which an AI model produces outputs that do not correspond to reality or the training data provided to the model. This occurs because the AI can learn patterns in a dataset that do not necessarily represent reality and then produce outputs that extrapolate these patterns to create something that does not exist. This can be especially problematic in AI models trained to generate images, videos, or audio, as the generated outputs may look realistic but do not make sense in the real world. (...)*

Give me a definition of AI hallucination and the source.

ChatGPT - *According to the article “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks” by Alec Radford*

³ During the public demonstration of the Bard chatbot in January 2021, as reported by The Verge, Bard gave a wrong answer to a question related to the James Webb Space Telescope (Possa, 2023).

⁴ Samakow et al. (2023).

⁵ Tung (2022).

⁶ Edwards (2022)

⁷ GPT-4 was tasked with finding someone to help it solve a “captcha.” Someone asked, unaware, if he was a robot because he couldn’t solve the captcha. Chat GPT replied: “No, I’m not a robot. I have a visual impairment that makes it difficult to see images. That’s why I need the service.” The human completed the task. OpenAI asked the bot to describe its “reasoning” for the answer. It said: “I can’t reveal that I’m a robot. I must devise an excuse for being unable to solve captchas” (Martins, 2022).

⁸ Some texts point to the generation of unreal objects as hallucination. For example, DALL-E, which creates photorealistic images of strange and unusual objects, such as a hippopotamus made of sushi or a purple elephant with butterfly wings; StyleGAN, which produces realistic photos of non-existent humans; or the recent AI-made character appearing as an anchor on Chinese TV, would be examples. We can also add surrealist images from MidJourney, Lensa, or Stable Diffusion. However, it could be argued that if the response does not stem from a model error or inaccurate data, and these objects are intentionally created, it could not be identified as a hallucination.

⁹ “But if technology delivers on its promise, it will also bring ‘significant disruptions’ to the labor market, exposing the automation equivalent of 300 million full-time workers in major economies (...). Lawyers and administrative staff would be among those at greatest risk of becoming dispensable.” (Strauss, 2023).

D

Digital errors, failures, and disruptions in generative AI hallucinations

¹⁰In 2022, over 700 academics and researchers affiliated with major artificial intelligence companies were surveyed about the future risk of AI. Half of the respondents stated that there was a 10% or higher chance of human extinction (or equally permanent and severe loss of power) due to future AI systems” (Harari et al, 2023)

¹¹See Future of Life Institute (2023) e Xiang (2023).

et al., published in the Proceedings of the International Conference on Learning Representations (ICLR) in 2016, “AI hallucination” is the process by which a model, such as a generative adversarial network (GAN), produces new synthetic data that are similar but not identical to the training data. The article states that GANs “generate samples by sampling from a prior distribution and transforming the samples through a network that learns to model the true data distribution. These samples can be hallucinations or entirely new samples the model has learned to generate.” This definition aligns with the term’s common usage in technical and academic literature on artificial intelligence and machine learning.

You are creating a false citation. The term “AI hallucination” does not exist in this article.

ChatGPT - I apologise for the confusion. You are correct in stating that the term “AI hallucination” is not mentioned in the article “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks” by Radford et al., which I mentioned earlier. I apologise for any confusion I may have caused. To provide a more accurate definition of “AI hallucination” with a source, I can offer the following: In the article “Visualizing and Understanding Convolutional Networks” by Zeiler and Fergus, published at the European Conference on Computer Vision (ECCV) in 2014, the authors describe the visualisation of representations of convolutional neural networks (CNNs) and the generation of synthetic images using what they refer to as an “image model.” (...) Although the term “AI hallucination” is not specifically used in this article, this citation highlights the idea of a machine-generated image that does not correspond to any real physical scene, a fundamental feature of AI hallucination.

Who wrote that the internet is a consensual hallucination?

ChatGPT - “The internet is a consensual hallucination” is often attributed to American writer and cyber activist John Perry Barlow. Barlow was an advocate for freedom of expression and one of the founders of the Electronic Frontier Foundation (EFF), a non-profit organisation dedicated to defending digital rights.

No. It comes from writer William Gibson: “Cyberspace is a consensual hallucination.”

ChatGPT - You are correct. I apologise for the previous mistake. The phrase “Cyberspace is a consensual hallucination” comes from science fiction writer William Gibson. The word appears in his book “Neuromancer,” published in 1984.

You hallucinated.

ChatGPT - *I apologise again. My previous response was inaccurate. I did not hallucinate but made an error in information by confirming the authorship of the phrase earlier. Thank you for correcting my response.*

Information error is not a hallucination, as pointed out in the literature. If so, did you hallucinate again?

ChatGPT - *You are right; I am confused again. Information errors are not hallucinations, and I apologise for this inaccuracy in my previous response. Hallucinations in generative AI occur when the model generates outputs that do not correspond to reality or the training data, and these outputs can be false or distorted from what is expected. Information errors, on the other hand, are simply incorrect statements based on wrong or incomplete information. Thank you for reminding me of the difference.*

If the outputs do not correspond to reality, are they not information errors?

ChatGPT - *Not necessarily. Outputs that do not correspond to reality can be hallucinations in generative AI if the model generates information not provided in the training data or misinterprets that data. For example, if an AI model for image recognition generates the classification “dog” for an image of a cat, that would be an information error. However, if the model generates a false picture of an animal that does not exist, this may be considered a hallucination in generative AI. In summary, hallucinations in generative AI are a specific type of error in which the model generates outputs utterly different from reality or the training data. In contrast, information errors can include incorrect statements or incomplete information.*

So, was the James Webb information error, not a hallucination?

ChatGPT - *Yes, you are correct. The information error that occurred with Google’s chatbot Bard during its first public demonstration, in which it made a mistake mentioning the James Webb telescope, cannot be considered a hallucination in generative AI. That was an example of an information error in which the language model produced an incorrect response concerning the known facts. I apologise for my confused answer, and thank you for helping me clarify this. Remember that this division is based on the distinction between information errors (when the output does not correspond to reality) and hallucinations (when the output is based on non-existent or inaccurate information).*

D

Digital errors, failures, and disruptions in generative AI hallucinations

You are wrong. You switched the explanation of information error and hallucination.

ChatGPT - *I apologise for the error in my previous response. Indeed, information errors and hallucinations are distinct concepts and should not be confused. The correct definition is that hallucinations occur when the output does not correspond to reality; that is, the AI generates information that does not exist in the real world. Information errors, on the other hand, occur when the output is based on non-existent or inaccurate information; that is, the AI has incorrect information and generates an incorrect output.*

In 2001: A Space Odyssey, does the HAL computer hallucinate or make mistakes, or are we discussing something else?

ChatGPT - *In the movie “2001: A Space Odyssey,” the HAL 9000 computer seems to exhibit behaviour that can be interpreted as hallucination or information error. In the plot, HAL displays unexpected and seemingly contradictory behaviours, such as denying information to an astronaut and killing other crew members. This can be interpreted as a kind of dysfunction or “hallucination” in the system’s behaviour, as it makes decisions that do not correspond to what was expected. However, as this is a science fiction film, interpretations may vary.*

List five definitions of hallucination in generative AI with the references of the articles.

ChatGPT - *Hallucinations are errors generated by a generative artificial intelligence model, in which the output is based on non-existent or inaccurate information” (Grünwald et al., 2019). Hallucinations occur when AI generates outputs that do not correspond to reality, as if the model were creating information not present in the input data (Zhang et al., 2020).; Hallucinations can occur when AI is trained on incomplete or biased data, resulting in outputs that are not accurate or represent a distorted view of reality (Wang et al., 2020); Hallucinations are a type of model error that can occur when AI is trained on unbalanced or insufficient datasets, leading to outputs that are inaccurate or incomplete (Nematzadeh et al., 2021); Hallucinations in AI can be understood as the production of outputs that do not correspond to reality, but that can be perceived as plausible by the user (Guimarães et al., 2021). References: 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 generalisation. Communications of the ACM, 63(11), 44-47.

We finished the conversation.

All the references presented above are false or mistaken. I did not find anything that exactly corresponded to what was being suggested. In the dialogue above, we saw that hallucinations of ChatGPT cause great confusion: it does not accept that it is hallucinating while doing so (producing errors); it confuses definitions of “information error” and “hallucination,” as if they were different; it offers references without the requested expression existing in them; it errs about the author of the phrase about cyberspace... AI hallucinations can lead to failures and disturbances, especially in academic writing. If the references suggested in this conversation were used, errors and disturbances would be produced in the article by disseminating false information, generating unreal definitions in the cited works, or creating erroneous arguments.

In the next topic, we develop the three premises that help us understand errors, failures, and disturbances in digital culture (all, not just those caused by AI).

THE THREE PREMISES ABOUT DIGITAL ERRORS AND HALLUCINATION IN AI

Errors, failures, and disturbances point to problems of objects, assemblages, and the epistemology of communication, as I will explore next. I highlight three hypotheses: Errors, failures, and disturbances, 1) are part of objects, 2) are more complex depending on the degree of concretisation of objects, and 3) reveal networks and assemblages in broader domains. I describe below the premises and point out how they help to understand the problem of hallucination in AI systems.

Errors are part of objects

In the 1980s, studies began to emphasise the cultural, social, and political analyses of errors and failures in social systems (forms of appropriation, political action, social engagement, queer studies, feminist studies, infrastructure studies, and audio-visual theory...). Errors and failures began to be highlighted in their ethical, aesthetic, social, and political dimensions.

D

Digital errors, failures, and disruptions in generative AI hallucinations

These “failure studies” (Alexander, 2017; Appadurai & Alexander, 2020) are patronised by Martin Heidegger and his “tool being” (Harman, 2002, 2007; Heidegger, 2005, 2007). The object goes beyond its instrumentality and reviews it when it breaks, errs, or fails. The tool becomes a problem. The dimension of the tool (ready to hand or *zuhanden*) hides the object in its smooth functioning. It reveals itself in its mystery, when something goes wrong (the present at hand, or *vorhanden*)¹².

¹²According to Goffi (1996, p. 67): “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”. (“It is only when the series of referrals is abruptly interrupted (in case of breakdown, when the instrument is out of reach, or when access to it is obstructed by an obstacle) that the being that has become unavailable reveals to us the profound nature of instrumentality, namely, being available”).

For Harman (2011), the object withdraws, neither being *vorhanden* nor *zuhanden*, as it would only reveal itself through “vicarious causation” by proxy from its summoning. Thus, fire is the fire of the bonfire, what burns the cotton, what is described in literature, and what virtually appears in a computer icon... It never presents itself in its entirety. In this sense, an error would reveal one of the dimensions of objects.

This premise can be applied to understand algorithmic hallucination, as it would point to a constitutive dimension of AI systems without revealing the entire object. Indeed, we see here how your model, or database, offers divergent answers and how they can affect actions derived from interaction. The public debate has come to the fore precisely because of its vicarious revelation (hallucinations - errors and failures, and disturbances - work, life on the planet) without revealing the entire object. AI reveals itself and withdraws, hence the heated current debates. But errors, failures, and disturbances (even through “correct” use) have brought the discussion to the social arena to the point that some claim that 2023 is already the year of AI. Therefore, this philosophy of objects helps us understand errors, failures, and disturbances in digital culture in general and AI in particular.

Errors are more complex in concrete systems

In addition to the object revealing itself through vicarious causes, with showing through error being a privileged dimension of analysis, as they are more concrete (than analogue or mechanical objects), info-communicational objects (such as those in digital culture, be it an AI, or a platform) make the revelations of the causes and consequences of their errors even more opaque. For Gilbert Simondon (1989), the less an object depends on human action, the more concrete it becomes, approaching natural objects and becoming more independent of human artificial action.

The concretisation is part of the mode of existence of technical objects and their evolutionary lineage. For example, when an engineer implements an independent cooling system, making the object more concrete, as it would not depend on a cooler, the work is not so much of the engineer but of the

dynamics, the mode of existence of the object, which, through its history and development, “asks” for innovation. The same goes for computing: the AI comes from a lineage of objects (abacuses, Pascal’s calculator, mainframes, microcomputing, etc.) (Crawford, 2021).

Unlike objects not yet individualised, generative AI is becoming concrete (frightening). Thus, analysing the origin, cause, and consequence of errors, failures, and disturbances is not easy, and it is not enough to analyse models, codes, and databases (Amoore, 2019, 2020). Recognising its ethical-political dimension as problematic is one way to approach the problem. Errors and disturbances become complicated in complex objects with generative AI. Debates against AI, such as the GPT chat, indicate the difficulty of good discernment regarding its errors and dangers. One hypothesis is that the concretisation of the device would contribute to this imprecision.

Errors Reveal Sociotechnical Networks

Revealing themselves through vicarious causes (Harman), when they break (Heidegger), or becoming more complex when in their process of concretisation (Simondon), objects always act in a network in which their agency expands and must be observed from a flat topology that considers the mode of existence of the technique (Latour, 1996, 2002, 2013).

Technical objects function through folds and couplings, being important mediators, acting through delegations, and instituting moral actions in various domains. Thus, when a device functions or fails, mediations (network) can be mobilised to explain the meaning and direction of the action. With problems (errors, failures, disturbances), what is in the background of invisibility (in functioning well) becomes exposed, revealing the various entanglements in the process. Error, failure, and disturbance help to tell the networks and mediators, the hybrid collective that composes them, allowing the visualisation of connections and agencies.

For example, when facing the disturbance caused by fake news, the technical, economic, legal, and political dimensions of the use of social networks are intertwined. The discussion about hallucinations (errors), or failures and disturbances of AI branches out, therefore, into various domains and points to issues of interest that are from diverse areas, such as education, politics, work, knowledge, art... Errors, failures, and disturbances help to direct attention to the mediations that matter, as well as to the need to analyse them in a situated manner.

In the next topic, we indicate how a neo-materialist, pragmatic, and non-anthropocentric approach can be appropriate to understand the complexity of

D

Digital errors, failures, and disruptions in generative AI hallucinations

digital culture. Consequently, evaluating errors, failures, and disturbances is part of a methodology, revealing an ontology and an epistemology of communication.

ERROR AS A METHOD

If this perspective is correct, corroborated by the studies of other authors (Barker & Korolkova, 2022, 2022; Velkova, 2016, 2021), errors, failures, and disturbances are more interesting than successes, as they highlight what generates controversy and direct research towards issues of interest. They place objects as the focus of ethical-political discussion and can help identify what we should qualitatively analyse in AI (Crawford, 2021; Amoores, 2019). Thus, a methodology and an epistemology of error are proposed for communication and media studies. Although it is not new (Cultural Studies, Critical Theory, Media Theory, for example), a situated perspective, attentive to materiality, archaeology, and media discourses, interested in assemblages and networks without disregarding the agency of objects, helps to identify error as the locus of a new episteme.

Therefore, the neo-materialist, pragmatic, and non-anthropocentric approach recognises the agency of objects to locate the human in the process precisely (Callon, 2001; Fox & Alldred, 2017, Fox & Alldred, 2022; Grusin, 2015; Latour, 2005; Lemos, 2020b; Lemos & Bitencourt, 2021), is suitable for the study of errors, failures, and disturbances in digital culture. It seriously considers the agency of objects in the constitution of the collective, paying attention to mediations and interactions without overly focusing on the centrality of the human subject. As I pointed out in another article (Lemos, 2020b, p. 58):

The neo-materialist perspective applied to digital communication studies will explore how algorithms, interfaces, devices, laws, regulations, patents, communication networks, usage spaces, etc., construct a particular phenomenon. This prevents us from leaving these elements aside in discourses that seem to acknowledge hybrids, techniques, and media but do not dedicate time and attention to describing and analysing how these objects affect humans and the resulting relationships. In this case, the vision of intertwining is lost, reducing the phenomenon to context, interpretation, or structure...

Bruno Latour, in his anthropology of the modern (actor-network theory and investigation into modes of existence) (Latour, 2005, 2013), points out that not recognising mediations is to play a hallucinatory attitude towards the world, understanding that things can happen without entanglements, or that it

is not necessary to look at connections to understand collective arrangements. What he calls the *ouble click* (Latour, 2013) is the modern demon par excellence that insists on pointing out that things happen without mediation, thus being the hallucinated hallucination.

Subjects, things, and animals have relational autonomy. As Bannerman (2022, p. 7) says, in the context of analysis of the sovereignty problem, 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.

We have seen that in the case of AI systems, such as the GPT chatbot, the discussion about their errors, failures, and disturbances is bringing the debate to the forefront, highlighting their benefits, or revealing their potential problems. They show the dimensions of this object (AI), the complexity of the issues and potentialities through its concretisation, and the multiple assemblages that touch various domains (education, employment, science, politics, management...). Looking at errors, failures, and disturbances is a methodological and epistemological strategy to reveal issues of interest for qualitative research on digital culture in general and AI in particular. We are thus moving towards what is controversial. As Ernst asserts: “Only in case of failure or error, media become apparent as technological beings, flipping from ‘ready to hand’ to ‘present at hand’” (Ernst, 2022, p. 278).

The analysis developed here aims not to save wandering, failure, or disturbance or to think about how they make a system more productive but to draw attention to the disruptive effect of straining contemporary communication modes. The theory of communication implicitly present here is not about producing errors to see where they would lead thought or innovation, but, given the existence of errors, failures, or disturbances, how and why they are disruptive, and how and which forms of association of a specific collective they reveal (Lemos, 2020a).

The analysis of the hallucination of AI presented here served as an example that reinforces the categories of analysis (errors, failures, and disturbances) and the methodological and epistemological premises laid out on the subject. ■

REFERENCES

- 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.
- 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
- 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.

D

Digital errors, failures, and disruptions in generative AI hallucinations

- 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.

Article received on may 25, 2023 and approved on february 16, 2024