Generative Artificial Intelligence and Undesirable Output: Developing a Taxonomy for Undesirable Output with a Focus on Hallucination

By Eric E. Cohen

Eric E. Cohen, CPA, is a technologist with a passion for collaboration toward the goal that "a piece of business information, once entered into any system, anywhere, never needs to be retyped as it moved through the business reporting supply chain." He's also a prolific author, engaged in virtually every effort to standardize accounting and audit data, a national expert to a wide variety of standards efforts, and co-founder of XBRL.

The year 2023 has brought with it the rise of generative artificial intelligence (gAI), bringing the chatbot to the masses. Different gAIs interact in different ways with users – some factual, reminding the user that it is just a "large language model" (LLMl, some programmed to act more like a human, with fun and familiarity, with a façade of a human with feelings, opinions, and interests.

When users feel like there is someone human on the other end, or make some other emotional attachment, they may trust the results provided by the gAI to a greater degree. That's why it can be jarring when those results are not the ones desired or expected. These undesirable results are often called "hallucinations."

The term "hallucination" is being used as one of the biggest concerns when using – or perhaps more directly, relying on the output of – a gAI tool based on an LLM. It may, however, be more precise and more helpful when considering how to fix, compensate for, or otherwise leverage the output from a gAI tool.

Chatbots based on LLMs generate content based on their internal models, developed through training, and probabilities, and not from "reality." There is no validation of input or output for suitability for specific uses. Thus, results may not be factually correct for numerous reasons: the desired output may not have been developed previously, or there is insufficient training data on the topic, or there has been incorrect or adversarial input – LLMs do not inherently possess "common sense."

> Chatbots based on LLMs generate content based on their internal models, developed through training, and probabilities, and not from "reality."

Even when gAI solutions, such as Bing AI Chat or Google Bard, incorporate search lookup or permit checking through search lookup, they are not reconciling the gAI output with the search output.

As such, responses may sound plausible but be factually incorrect. Citations and references are not drawn from a database, nor are they verified, without some manner of extension or plug in.

However, the term "hallucination" may be used for a broader group of undesirable outputs than it should get credit for. Therefore, it may be helpful to develop a taxonomy of gAI undesirable output results, with a focus on hallucinations, as we will describe them.

Brief History

Chatbots are not a new phenomenon, with a start date of November 30, 2022 (when OpenAI delivered ChatGPT to the world and started the generative AI tidal wave). Eliza (1966) was the name of an early gAI chatbot out of MIT, a recent Peabody award winner,[1] often called the world's first autonomous computer chat program. There are accounts[2] that the secretary of the scientist who developed the code was one of many who treated Eliza as a confidant. That scientist, Joseph Weizenbaum, went on to write on AI and the need to not rely on it for important decisions, as they lack compassion and wisdom. In his book *Computer Power and Human Reason*,[3] he made a distinction between "deciding" and "choosing" – deciding is a computation activity, where choosing comes from judgment, wisdom and compassion, and not calculation. He concludes. "What emerges as the most elementary insight is that, since we do not now have any ways of making computers wise, we ought not now to give computers tasks that demand wisdom."

Your author has been using chatbots for more than 30 years to help brainstorm and organize his thoughts. Dan E. Burns was the developer of *Thoughtline* (MS DOS) and *Write-EZ* (Windows) creativity tools, patterned after Eliza. The author wrote of this phenomenon 20 years ago, and interviewed Mr. Burns about the usefulness of the technology. Burns' words seem as relevant today as in 2003:

"From experience, I know that being an entrepreneur and owning or managing a small business can be a lonely experience. Those managers are thinking through their communications on their own. Thinking is not a monologue, but an internal, condensed dialogue. WriteEZ provides a companion for that dialogue, so you can advance your thinking. WriteEZ makes that process explicit and easier by asking the questions for you."

Burns' design was based on "directed discussion" – "a heuristic technique for helping students develop essays by asking questions, challenging assumptions, requiring supporting evidence, and then critiquing a final draft."[4]

Fast forward. gAI tools became somewhat commonplace leading up to 2022. Most of the generation was images, or art. Other than IP concerns, people did not expect fact. Then came ChatGPT. A number of more obvious kinds of undesirable results emerged.

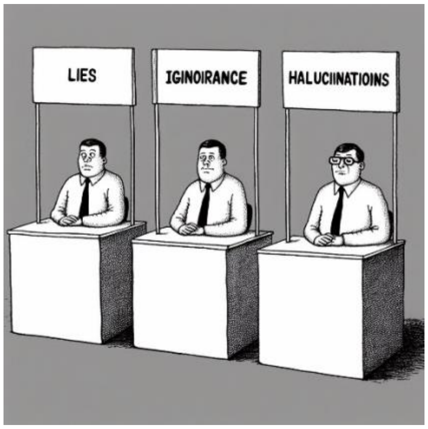

Figure 1: Created with DALL-E-4 from Bing AI Chat

Many guardrails are in place to deal with outputs that are offensive, hateful, violent or otherwise harmful.

- **Refusal to respond**: These are when chatbots say they cannot*. For example, a chatbot might say the question violated a guideline, was outside of the scope of an LLM, or that it didn't want to respond.*
- **Facts that are not facts**: These are when *chatbots* make up facts or details that are not true or supported by evidence*. For example, a chatbot might invent a research paper that does not exist or cite a source that is not reliable. However, some users would push back on correct responses to force incorrect responses to get these results.*
- **Semantic issues***: These are when chatbots produce content that is grammatically correct but does not make sense or is irrelevant to the context. *For example, a chatbot might switch topics randomly, or use words that are out of place.* A counterpart might be as illustrated in Figure 1; DALL-E-4 sometimes refuses to provide the text it is given when creating images.
- **Stylistic issues**: These are when chatbots generate content that does not match the tone, genre or purpose of the task. *For example, a chatbot might use informal language in a formal setting, or mix up different styles of writing*, harmful or inappropriate for the situation. *For example, a chatbot might use hate speech, promote violence, or violate privacy.* These are often manipulation by "malicious" users. Some may feel that harmful or inappropriate outputs should be broken out as particularly egregious; many guardrails are in place to deal with these outputs that are offensive, hateful, violent or otherwise harmful. They may be due to malicious users or to problems with the model's understanding of social norms and ethical guidelines.

Hallucinations

The author has experienced many confident responses from chatbots that are not true. These include:

1. The results of a request for the biography of *ThinkTwenty20*'s managing editor.
2. The lyrics for the classic 1967 hit by Vaughn Horton, *Pennsylvania Turnpike, I Love You.*

3. A list of articles by leading scholars from reputable academic journals on AI in accounting.

We know the news about a New York attorney[5] who included citations for his court filing that were completely fabricated by ChatGPT. (But why wasn't he using a real citator?)

In each case, the AI delivered a complete and confident reply that was completely wrong. It wrote a completely new song about Pennsylvania's Turnpike. It took real journal names, real (if mixed up) author names, and created plausible sounding article titles that simply didn't exist.

I do not call these hallucinations – I call them math. Others disagree. But whether hallucination or not, these results can have negative consequences for the reliability, credibility and trustworthiness of gAI applications and can potentially harm users of society.

The ISO does not currently have a definition of "hallucinations" when it comes to artificial intelligence, although it has quite a few definitions of "artificial intelligence" and a helpful technical report on testing AI (see Additional References and Glossary). Therefore, different groups can include or exclude different categories of undesirable results as hallucinations or not.

(With the example of the song above, it may not have been the result I wanted, but it could have been the result I needed – it invented a song! One person's undesirable result is another person's treasure.)

A hallucination implies that the chatbot is generating content that is not influenced by external inputs, but rather by its own internal logic and probabilities.

In a recent survey on hallucination in natural language generation,[6] the authors credit the genesis of the term "hallucination" to *Computer Vision,*[7] with positive meaning, but later took on more negative implications. The authors took the position that hallucination is both generated text that is *unfaithful to the provided source input* (with which the author agrees) and *nonsensical* (which the author finds troublesome, as an AI does not possess "sense"; it decides computationally, it does not choose, as Weizenbaum noted.). They do helpfully categorize hallucinations as *intrinsic* or *extrinsic,* based on whether the undesired result *disagrees with* or *cannot be verified from* source content.

Of these, the author places hallucinations largely in the Semantic issues arena. A hallucination implies that the chatbot is generating content that is not influenced by external inputs, but rather by its own internal logic and probabilities. Reasons may include "drift", "degradation" and "staleness"[8]; Bing AI Chat limits the back and forth on a topic for these reasons.

A Bard query provides definitions to these terms, and the relevance to our topic:

- **Drift** refers to the tendency of large language models to generate text that is less relevant to the topic at hand over time. This can happen because the model is not able to keep track of the context of the conversation or because it is simply generating text based on the most likely next word in the sequence.

- **Degradation** refers to the tendency of large language models to generate text that is of lower quality over time. This can happen because the model is not able to learn from its mistakes or because it is simply being asked to generate text on topics that it is not familiar with.
- **Staleness** refers to the tendency of large language models to generate text that is outdated or inaccurate. This can happen because the model is trained on a dataset of text that is not up-to-date or because it is simply not able to keep up with the rapidly changing world around it.

Bard also suggests that “overfitting” (see glossary) can be a problem, where the model fits the training data too well and cannot generalize in response to use input accurately.

As to the other examples above, your author agrees they are undesirable results, but not hallucinations.

Take Tay, for example. The Microsoft 2016 AI experiment gone bad was not hallucinating, but rather repeating what it learned from others. Nevertheless, Tay still illustrates some of the ethical and social challenges of creating chatbots that can learn from human interactions.

Likewise, concerns about “bias” are not related to lack of faithfulness to inputs, but the nature of the training data itself.

That training data may be an issue, where there is a collapse of the AI model itself due to the model entering a feedback loop – where the AI is trained on data it generated itself, sometimes called Model Autophagy Disorder (MAD).[9] This is a different kind of “bias,” an AI “echo chamber.”

Users should be aware that generative AI systems are not human and that they do not possess the same abilities as humans.

Developing a Taxonomy

With all of the being noted, one can begin to build a “taxonomy,” an organizational structure that may be helpful when considering how to remediate the risks of reliance on AI. We have discussed:

- Cause (e.g., drift, degradation, staleness, model issues)
- Focus (e.g., intrinsic, extrinsic)
- Classification (e.g., factual, semantic, stylistic)
- Operational (e.g., guardrails, refusal)
- Expectations (e.g., recognition it is computational, not a database/sense/choice)
- Planned uses (e.g., casual use, business communication, mission-critical operations)

- Known risks (e.g., IP violations, incorrect/biased training materials)

Note: Faithfulness can lead to undesirable results; nonsense always leads to undesirable results. A classification, or taxonomy, of undesirable output results into factual, semantic, stylistic and harmful or inappropriate categories can be helpful. Such tools can be used to develop targeted mitigation strategies. For example, to mitigate factual errors, we can develop techniques to improve the AI's ability to verify the accuracy of its output. To mitigate semantic errors, we can develop techniques to improve the AI's understanding of language and its ability to generate text that is coherent and relevant. To mitigate stylistic errors, we can develop techniques to improve the AI's understanding of different writing styles and its ability to generate text that is appropriate for the situation. To mitigate harmful or inappropriate outputs, we can develop techniques to identify and filter out harmful or inappropriate text.

Operational aspects of undesirable output results are important. For example, we need to consider how to detect undesirable output results in real time and how to respond to them. We also need to consider how to communicate the risks of undesirable output results to users. User expectations are also important, but for such a rapidly evolving environment, education is very difficult. For example, users should be aware that generative AI systems are still under development and that they may produce undesirable output results. Users should also be aware that generative AI systems are not human and that they do not possess the same abilities as humans.

Figure 2: Another image from DALL-E-4

Note: Claude (Claude.ai) recommends:
To develop the taxonomy further:

- Categories could include the type (factual, semantic, etc.), cause (e.g., lack of context, overconfidence), expected user impact (minor confusion to harm).
- Subcategories could distinguish intrinsic vs. extrinsic outputs, or examine the content (made up details, switching topics) vs non-responses.

- Causal factors, such as data biases, model deficiencies and human manipulation could be noted.
- Mitigation approaches could address the AI system design (e.g., uncertainty metrics), data & models (e.g., debiasing) and user education.
- Case examples that don't neatly fit proposed categories could help refine the taxonomy definitions.
- Real-world use scenarios could inform what level of unwanted output is acceptable.

Application

Let’s take one of our examples from above: the author asks the gAI to provide a bio for the managing editor (to be used in an upcoming presentation). The AI provides a detailed description that has no connection to the managing editor’s experience.

- Cause: While not a desirable result, the LLM is not a search engine. The cause is math.
- Focus: While we can’t ask the LLM to tie back to the results, there was likely similar content in the training.
- Classification: Factual**.**
- Operational: There are no obvious operational issues.
- Expectations: We were looking for facts as if it were a database.
- Planned uses: Business communication.
- Known risks: Potential embarrassment, false claims, missing the good stuff that might influence others.

Conclusion

While more narrowly defining “hallucinations” may seem like nitpicking when generative AI is creating nonsense, presenting incorrect information as fact with great confidence, and potentially influencing users that treat the AI as an extension of themselves, helping users understand why gAI may provide unreliable and undesirable results as well as working toward plans to remediate and mitigate.

There are many dimensions to these problems. It ranges from the users themselves (decoupling them from the experience, turning on appropriate skepticism) to improving the training data, models, interfaces, guardrails and other logic.

* * *

Additional References:

ISO/IEC TR 29119-11:2020(en) Software and systems engineering — Software testing — Part 11: Guidelines on the testing of AI-based systems.

ISO/IEC 23053:2022(en) Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML).

Glossary (Source is the ISO Online Browsing Platform at https://www.iso.org/obp/ui#search unless otherwise noted).

Artificial Intelligence –[Note: Oddly enough, defined numerous times]

ISO/IEC 2382:2015 *capability of a functional unit to perform functions that are generally associated with human intelligence such as reasoning and learning*

ISO/IEC 39794-16:2021: *discipline concerned with the building of computer systems that perform tasks requiring intelligence when performed by humans*

ISO 22458:2022: *engineered system with capability to acquire, process and apply knowledge and skills*

Overfitting

ISO/IEC 23053:2022(en) <machine learning> creating a model which fits the training data too precisely and fails to generalize on new data.

AI Process, as told by ChatGPT

The following is a response by ChatGPT when asked about the processes in ChatGPT. Risks and errors can be introduced at each step. The steps may be summarized in the following graphic:

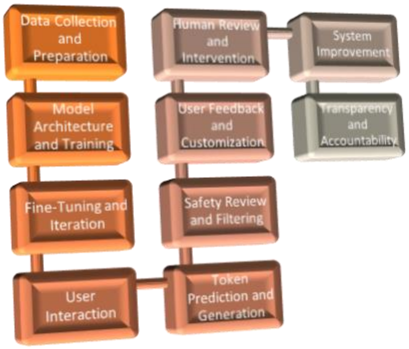
Data Collection and Preparation
Model Architecture and Training
Fine-Tuning and Iteration
User Interaction
Token Prediction and Generation
Safety Review and Filtering
User Feedback and Customization
Human Review and Intervention
System Improvement
Transparency and Accountability

Figure 3: Steps of gAI Process from ChatGPT

ChatGPT's response begins here:

"I can provide you with a generalized depiction that can help you understand the high-level process involved in creating and using generative AI systems:

1. **Data Collection and Preparation:**
    1. Gather a diverse and relevant dataset for training the AI model.
    2. Clean and preprocess the data to remove noise and inconsistencies.
2. **Model Architecture and Training:**
    1. Choose a suitable architecture (e.g., transformer) for the generative AI model.
    2. Train the model on the prepared dataset using techniques like supervised or reinforcement learning.
    3. Incorporate safety measures, guidelines, and constraints into the training process.
3. **Fine-Tuning and Iteration:**
    1. Fine-tune the model based on specific use cases or user feedback.
    2. Continuously iterate and improve the model's performance and safety.
4. **User Interaction:**

    1. Users provide prompts or input to the generative AI system.
    2. The input is tokenized and processed by the model.
5. **Token Prediction and Generation:**
    1. The model predicts the next tokens in the sequence based on the input and its training.
    2. Generated tokens form a coherent response.
6. **Safety Review and Filtering:**
    1. The generated response undergoes safety review, including checking for offensive, harmful, or inappropriate content.
    2. Real-time filtering tools can be used to catch problematic outputs.
7. **User Feedback and Customization:**
    1. Users provide feedback on the generated output.
    2. Some systems allow users to customize the behavior of the AI within certain bounds.
8. **Human Review and Intervention:**
    1. In cases of uncertainty or nuanced content, human reviewers assess the generated response.
    2. Human reviewers help maintain the quality and safety of the system's outputs.
9. **System Improvement:**
    1. Feedback from users and human reviewers contributes to improving the system's safety measures and performance.
    2. Developers refine the model and update its training based on user needs.
10. **Transparency and Accountability:**
    1. Developers maintain transparency about the AI's behavior, safety mechanisms, and guidelines.
    2. Users are informed about how the AI system operates.

- Please keep in mind that the specifics of this process can vary based on the organization, technology, and the type of generative AI system."

ChatGPT's response ends here.

Endnotes:

[1] ELIZA wins Peabody Award | MIT CSAIL.

[2] 'Please Tell Me Your Problem': Remembering ELIZA, the Pioneering '60s Chatbot (mentalfloss.com).

[3] Computer power and human reason : from judgment to calculation : Weizenbaum, Joseph : Free Download, Borrow, and Streaming : Internet Archive.

[4] A history of the development of the software is documented in *Writing Teachers Writing* Software, taken from jhttps://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=91e63cec4f60fad9bfdedeaca52b237e9fd7e7b9 p 72-75.

[5] ChatGPT is not quite ready to be your lawyer (morningbrew.com).

[6] https://dl.acm.org/doi/abs/10.1145/3571730, https://arxiv.org/pdf/2202.03629.pdf.

[7] S. Baker and T. Kanade. 2000. "Hallucinating Faces." In Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580). 83–88. https://doi.org/10.1109/AFGR.2000.840616

[8] https://www.iso.org/obp/ui#iso:std:iso-iec:tr:29119:-11.

[9] https://arxiv.org/abs/2307.01850.