# BEWARE OF BOTSHIT: HOW TO MANAGE THE EPISTEMIC RISKS OF GENERATIVE CHATBOTS

**Timothy R. Hannigan**
University of Alberta
Alberta School of Business
11203 Saskatchewan Drive NW
Edmonton, AB
T6G 2R6
Canada

**Ian P. McCarthy**
**(corresponding author: ian_mccarthy@sfu.ca)**
Simon Fraser University,
500 Granville St., Vancouver, BC,
V6C 1W6
Canada

and

Luiss
Viale Romania, 32
00197 Roma
Italy

**André Spicer**
Bayes Business School
City University of London,
106 Bunhill Row
London
EC1Y 8TZ
United Kingdom

*Forthcoming in Business Horizons*

# BEWARE OF BOTSHIT: HOW TO MANAGE THE EPISTEMIC RISKS OF GENERATIVE CHATBOTS

## ABSTRACT

Advances in large language model (LLM) technology enable chatbots to generate and analyze content for our work. Generative chatbots do this work by ‘predicting’ responses rather than ‘knowing’ the meaning of their responses. This means chatbots can produce coherent sounding but inaccurate or fabricated content, referred to as ‘hallucinations’. When humans *use* this untruthful content for tasks, it becomes what we call ‘botshit’. This article focuses on how to use chatbots for content generation work while mitigating the epistemic (i.e., the process of producing knowledge) risks associated with botshit. Drawing on risk management research, we introduce a typology framework that orients how chatbots can be used based on two dimensions: response veracity verifiability, and response veracity importance. The framework identifies four modes of chatbot work (*authenticated*, *autonomous*, *automated*, and *augmented*) with a botshit related risk (*ignorance*, *miscalibration*, *routinization,* and *black boxing*). We describe and illustrate each mode and offer advice to help chatbot users guard against the botshit risks that come with each mode.

**Keywords**: chatbots, bullshit, botshit, hallucinations, large language models, artificial intelligence, epistemic risks

## 1.0 WHO’S A PRETTY POLLY

On November 20, 2022, OpenAI released its chatbot, Chat Generative Pre-trained Transformer (ChatGPT), for public use. This AI-driven chatbot generates text responses to human prompts, such as questions and requests. ChatGPT can undertake plagiarism checks, create content such as proposals, stories, applications, reviews, and jokes, and perform programming work such as creating and debugging code. In January 2023, Salesforce CEO Marc Benioff announced, “Just promoted #ChatGPT to our management team. It’s an invaluable asset, enhancing decision-making efficiency and injecting humor into meetings!” (Benioff, 2023). These endorsements are mirrored by studies showing that ChatGPT significantly improved college-educated professionals’ writing productivity and quality (Noy & Zhang, 2023). All of this excitement helped ChatGPT become the fastest-growing consumer application ever, achieving 100 million users within two months (Reuters, 2023).

Accompanying the rapid ascent of ChatGPT are concerns regarding our blind reliance on its accuracy. Stack Overflow, for instance, banned contributions from ChatGPT due to its frequent inaccuracies (Vincent, 2022). In response, Stack Overflow introduced its own generative AI tool, OverflowAI, designed to offer developers more dependable support (Chandrasekar, 2023). Worldwide, educational institutions have reevaluated their research methodologies (Lindebaum & Fleming, 2023), approaches to learning assessment, asking if ChatGPT is a "bullshit spewer or the end of traditional assessments in higher education?" (Rudolph, Tan & Tan, 2023: 1). In journalism, researchers guided by ChatGPT sought access to articles published in The Guardian newspaper. However, The Guardian couldn't locate these articles because they had never been written (The Guardian, 2023). Likewise, a Federal District Court of New York, sanctioned and fined for two lawyers for submitting a legal brief containing fictitious cases and citations, all generated by ChatGPT (New York Times, 2023). These content veracity issues extend to ChatGPT competitors. The professional services firm KPMG filed a complaint in the Australian Senate after a committee admitted to not verifying the accuracy of a document generated by Google's chatbot, Bard. The chatbot falsely accused companies of involvement in non-existent scandals and referenced individuals who had never been employed by those firms (Belot, 2023).

These limitations in the veracity of LLM-based chatbot technology have raised doubts about their reliability. With the convenience of using chatbots for efficient content generation, there's a concern that this technology will reduce the cost it takes humans to bullshit to zero while not lowering the cost of producing truthful or accurate knowledge (Klein, 2023). Computational linguists highlight this problem by explaining how LLMs are great at mimicry and bad at facts, making them a beguiling and amoral technology for bullshitting (Weil, 2023). Even in areas where the technology showed accuracy, such as in identifying prime numbers, researchers are finding that chatbots are increasingly prone to generating inaccurate answers (Chen et al., 2023). Regardless of whether this unreliability is a function of chatbot getting "lazy" (Mollick, 2023), or the underlying technology undergoing architectural shifts, the result is that chatbots there are doubts about the accuracy chatbot generate content.

OpenAI (2023), the company behind ChatGPT, acknowledges this risk of producing plausible but incorrect or nonsensical answers due to the absence of a source of truth during training. Also, studies assessing chatbot veracity show that in certain areas, like translation and news summarization, they perform comparably to human or commercial tools (Jiao et al., 2023; Zhang et al., 2023). However, these assessments typically focus on rigid "memorization" of structured data, lacking generative responses involving complex, unstructured data (Mukherjee & Chang, 2023; Zhao et al., 2023). This challenge of assessing and improving the factual accuracy of chatbots is so important that OpenAI promoted that its ChatGPT-4 is 40% more likely to produce factual content than ChatGPT-3.5 (OpenAI, 2023).

In this article, we explore how human users of chatbots (i.e., managers and other professionals) can avoid the dangers of using chatbot untruths for work. Thus, we focus on the text-based content generation capabilities of chatbots, and not their analysis, programming, and audio-image generation capabilities. We begin by examining the evolution of chatbots and the LLM technology underlying their potential to hallucinate untruths that humans use and transform into botshit (Table 1). We highlight that LLMs are designed to 'predict' responses rather than 'know' the meaning of these responses. LLMs are likened to 'stochastic parrots' (Bender et al., 2021) as they excel at regurgitating learned content without comprehending context or significance. LLMs rely on pattern analysis to predict suitable responses based on their training data but lack inherent knowledge systems, like the scientific method, to evaluate truthfulness. Furthermore, LLMs sometimes hallucinate by generating responses unsupported by their training data as they prioritize generating the best possible guess in their responses. (Xiao & Wang, 2021; Ji et al., 2023).

We use these insights into LLMs to distinguish between bullshit, hallucinations, and botshit and their implications (see: Table 2). Then drawing upon risk management research, we offer a framework (see Figure 1) to guide the effective use of chatbots for various work tasks while mitigating the epistemic risks of botshit. The framework assesses the importance and verifiability of response veracity, leading to four chatbot work modes (*authenticated*, *autonomous*, *automated*, and *augmented*), each associated with a botshit related risk (*ignorance*, *miscalibration*, *routinization*, and *black boxing*). We discuss practices for each mode's veracity-verifiability conditions and the related epistemic risks. We then offer guardrail related advice for how the technology, organizations and human users can mitigate these risks.

## 2.0 CHATBOTS UNVEILED: KNOWING VERSUS PREDICTING

A chatbot is a computer program that facilitates a natural dialogue (text or voice-based) with humans. Created in 1966, one of the first chatbots, ELIZA, was a mock 'Rogerian' psychotherapist (Weizenbaum, 1966). Rogerian psychotherapy involves the therapist reassuringly repeating a patient's statements back to them, something that this pioneering chatbot could do with early text processing technology. Many users of ELIZA reportedly felt they were talking to a real person, despite its creator's insistence that it had no real understanding of the conversations it participated in (Shum et al., 2018). In 1995, the chatbot ALICE (Artificial Linguistic Internet Computer Entity) used Artificial Intelligence Markup Language (AIML) to specify heuristic conversation rules and won the Loebner Prize in 2000, 2001, and 2004, a competition for computer programs considered to be the most human-like (Wallace, 2009).

Today's chatbot technology stems from a sub-field of computer science known as "natural language processing" (NLP). This sub-field of artificial intelligence (AI) aims to create technology that can understand, interpret, and generate human language in a valuable way (Kietzmann & Pitt, 2020; Przegalinska, 2019). OpenAI's ChatGPT, Google's Bard (powered by its Gemini Pro system), Anthropic's Claude 2.1, and X's (formerly Twitter) Grok, are chatbots based on an approach within NLP known as 'generative AI' (i.e., the G in ChatGPT). A generative chatbot uses neural networks to learn the semantic distance between words using vector coordinates assigned to the words. For instance, with English language training data, the word 'pasta' follows the word 'eat' more often than the word 'chaos, and in a vector space of words 'pasta' and 'eat' would be closer to each other than 'chaos'. This is how chatbots generate new human-like text-based responses to the prompts they receive.

The generative feature of chatbots involves 'pre-training' (i.e., the P in ChatGPT), then fine-tuning, and then users interacting with it. Users submit prompts (i.e., a set of instructions) to a chatbot that directs and enhances the response capabilities of its processing model. Generative chatbots produce natural language responses to prompts using LLMs, consisting of a large data set and an AI 'transformer' (i.e., the T in ChatGPT) trained using the large data set. The transformer is a neural network, highly interconnected neurons that work in unison on image and speech recognition and natural language processing (Kietzmann et al., 2020).

While LLM chatbots are powerful content generation and analysis tools, they do not have any sense of truth or reality beyond the words that tend to co-occur in their training data and processes (i.e., the supervised fine-tuning, the reward model, and policy optimization, see Table 1). This training data is often date-limited and necessarily retrospective. For instance, as of December 10, 2023, the current version of Open-AI's chatbot (i.e., ChatGPT-4) was trained on data up to April 2023. There are also copyright concerns about the training data. For example, a group of more than 8,000 authors, including Margaret Atwood, Jennifer Egan, George Saunders, and Jodi Picoult, recently wrote an open letter to generative AI companies asking for fair compensation on the use of their copyrighted materials (Brown, 2023). Also in California there is a class-action lawsuit against OpenAI for supposedly stealing personal data used in the ChatGPT training (Mauran, 2023). Although OpenAI has signed licensing deals with the Associated Press for access to their archives (O'Brien, 2023), there is looming uncertainty over what data the company will continue to use for training its LLMs.

For organizations using LLM chatbots, it's crucial to understand that they still need substantial human input to function effectively. Despite worries about LLMs taking over jobs (Chui, Roberts & Yee, 2022) and posing challenges to our understanding and knowledge acquisition, like AI in general, chatbots are tools that can often enhance, not replace, human tasks (Jarrahi, 2018; Jarrahi et al., 2023; Paschen, Pitt & Kietzmann, 2020). For example, Glaser and Gehman (2023) proposed how ChatGPT could help academic work as a research assistant, data analyst, and co-author. This use of the technology would be a conjoined agency approach (Murray et al., 2021) where academics use the Chatbots with a careful concern for precision and truth. For work in organizations in general, we suggest a similar and contingency-based perspective where the importance of chatbot response veracity and verifiability varies depending on the nature of the work.

To use LLMs responsibly, we need to know how to mitigate the epistemic risks that these technologies introduce. Epistemic risk is the likelihood that one's claims accurately represent the world (Babic, 2019). Therefore, it is up to organizations and humans who use and, in turn, train these chatbots to be aware of how they work and use this understanding to help mitigate the epistemic risk of using this technology. To help with this understanding, we draw upon research by OpenAI researchers (Ouyang, et al., 2022) to explain the key steps in how a LLM chatbot works (see Table 1). For each step, a concise and accessible description of what happens is given, along with an outline of the risks producing untruths that become applied by users.

Table 1 highlights that data accuracy, preprocessing, tokenization, and unsupervised learning (Steps 1-4) influence whether an LLM such as ChatGPT generates misleading content (hallucinations) or not. Ideally, LLMs would have data that are the basis for "ground truths" i.e., actual, definitive, and accurate data against which the LLM's predictions or outputs are compared (Munn, Magee & Arora, 2023). For example, a user asks a chatbot, "Which professor is known for coining the term 'open innovation'?" and the chatbot responds with "Professor Michael Porter," this response would be a complete departure from the ground truth ("Professor Henry Chesbrough" is the ground truth, see Chesbrough, 2003). However, if the prompt was "Which company is currently the leading practitioner of open innovation?", the truth is subjective and changing relative to some survey source that estimates company open innovation practices. These simple examples show that during steps 1 – 4 in Table 1, the relationship between input data and what constitutes ground truth can sometimes be clear-cut and sometimes it can be relative and subjective. This distinction also points to a more significant issue in the organizational literature on AI, which examines the uncertainty around ground truth labels used to train models (Lebovitz, Levina, & Lifshitz-Assaf, 2021).

Creating an LLM model grounded in training data is a multistep process. The data fed into an LLM must be carefully 'cooked' by a team of humans as it is akin to gathering and deciding what ingredients to include in a recipe. The cooks are a team of human renderers (Hannigan et al., 2019) who compile a data set (or corpus in NLP terms), and their decisions determine how reality and truth are to be conceptualized by the LLM (Kozyrkov, 2020). This means ground truth is determined by the cooks' expertise and biases and the company's business model employing them to feed the LLM. Thus, the adage 'garbage in, garbage out' highlights that LLM responses strongly depend on the credentials of the cooks and the quality of the cooked data used in steps 5-7. The company OpenAI that runs ChatGPT has a specific process called "Reinforcement Learning from Human Feedback" (RLHF) (Ouyang et al., 2022) that demonstrates these additional steps.

| **Reinforcement Learning from Human Feedback (RLHF): The ChatGPT LLM process** | **Description** | **Risk of generating LLM hallucinations** |
|---|---|---|
| 1. Data collection | A large text data set is compiled to capture diverse topics, contexts, and linguistic styles. | If the data is biased, not current, incomplete, or inaccurate, the LLM and human users can learn and perpetuate this its responses. |
| 2. Data preprocessing | The data is cleaned to remove irrelevant text and correct errors and then converted for uniform encoding. | Preprocessing inadvertently removes meaningful content or adds errors that alter the context or meaning of some text. |
| 3. Tokenization | The data is split into 'tokens', which can be as short as one character or as long as one word. | When language contexts are poorly understood, tokenization results in wrong or reduced meaning, interpretation errors, and false outputs. |
| 4. Unsupervised learning to form a baseline model | The tokenized data trains the LLM transformer to make predictions. The LLM learns from the data's inherent structure without supervision. | The LLM learns to predict content but does not understand its meaning, leading it to generate outputs that sound plausible but are incorrect or nonsensical. |
| 5. Reinforcement Learning from Human Feedback:<br>i) supervised fine-tuning of model (SFT) | A team of human labelers curates a small set of demonstration data. They select a set of prompts and write down expected outputs for each (i.e., desired output behavior). This is used to fine-tune the model with supervised learning. | This process is very costly, and the amount of data used is small (about 12,000 data points). Prompts are sampled from user requests (from old models). This means the SFT only covers a relatively small set of possibilities. |
| 6. Reinforcement Learning from Human Feedback:<br>ii) training a reward model (RW) | The human labelers repeatedly run these prompts against the SFT model and get multiple outputs per prompt. They rank the prompts for mimicking human preferences. This is used to train a reward model (RM). | Human labelers agree to a set of common guidelines they will follow. There is no accountability for this, which can skew the reward model. |
| 7. Reinforcement Learning from Human Feedback:<br>iii) fine-tuning SFT model through proximal policy optimization (PPO) | A reinforcement learning process is continually run using the proximal policy optimization (PPO) algorithm on both the SFT and RM. The PPO uses a "value function" to calculate the difference between expected and current outputs. | If faced with a prompt about a fact not covered by the training data (SFT and RM), the LLM will likely generate an incorrect or made-up response. |

Table 1: Why Large Language Models can be Full of It.

In Steps 5-7, prompts guide an LLM to enforce rules, automate processes, and ensure specific response qualities and quantities (White et al., 2023). These steps are crucial for fine-tuning the LLM to help address any misalignment problems from Steps 1-4, where outputs match training metrics but do not align with human users' values. Ramponi (2022) suggests that LLM misalignment issues include: i) unhelpful responses, ii) generating incorrect or fictional information, iii) outputs that are hard to interpret, and iv) producing biased or toxic content. With a misaligned baseline model, a vague prompt can exacerbate the risk of users receiving and using hallucinations and thus transforming them into botshit. This can occur by not properly setting the context for the response and the desired size and style the response should have. However, even clear prompts can lead to hallucinations and botshit if the LLM lacks data to provide a comprehensive, accurate response, potentially resulting in repetition or falsehoods.

Step 5 is the supervised fine-tuning (SFT) step of RLHF process. It aims to improve the LLM outputs to align the LLM around human preferences. Given the massive scale of the training data, it is not feasible to have humans check every possible prompt output. Instead, a team of human labelers select a small set of prompts from an older version of the LLM and specify the expected outputs for each prompt. The aim here is to guide the LLM towards more desired output behavior. This step will involve around 12,000 prompts (Ramponi, 2022), which is large and costly for a team of humans to handle and yet only covers a tiny set of possibilities for the LLM. This set of outputs is used to fine-tune the original LLM as part of a supervised learning process.

Step 6 of the RLHF process is an attempt to try to mimic human preferences. At this point, the LLM can still vary in subsequent outputs to the same prompts. Human labelers determine a set of common guidelines to rank multiple outputs per prompt for quality. These rankings are then used to generate a version of the LLM known as the reward model (RM). The final step (7) in the RLHF process is to use a proximal policy optimization (PPO) algorithm to combine insights from the SFT and RM models and feed them back into the production-ready LLM. With ChatGPT, new versions of the LLM are launched approximately once per year (i.e., GPT-3, GPT-3.5, GPT-4, and possibly soon GPT-5). Within each version, OpenAI continually runs RLHF steps 6 and 7 to optimize ChatGPT for human preferences better.

In sum, the steps in Table 1 show that generative chatbots are not concerned with intelligent knowing but with prediction (Agrawal, Gans & Goldfarb, 2018). While LLMs can be trained to predict content that will be useful and credible to carry out a work task as specified a prompt, the predicted response does not involve intelligent context-based knowing and decision-making. Instead, it generates a technical word-salad on patterns of words in training data (which is itself a black box). This means chatbots are machines that excel at predicting how to 'make stuff up' to prompts, which sometimes turn out to be correct. For example, using the TruthfulQA benchmark four LLMs were tested on 38 subjects like health and politics, using 817 questions (Lin, Hilton & Evans, 2021). The best-performing LLM was only truthful in 58% of cases, compared to a 94% human accuracy rate. This capacity to efficiently produce content that could be hallucinatory means chatbot users and their organizations face significant epistemic risks when using this technology for work. As a result, and as we explain later, users should consider response veracity verifiability and response veracity importance to avoid applying and transforming any potential hallucinatory content into botshit.

## 3.0 BULLSHIT AND BOTSHIT

We now draw on the growing research literature on bullshit to help understand that when humans use and apply chatbot content contaminated with hallucinations, this results in botshit (see Tabel 2). These insights guide the ideas for our framework on how to use chatbots for work, while mitigating the limitations and epistemic risks associated with a chatbot's inherent inability to understand meaning and generate truth claims.

Ever since the work by Harry Frankfurt (2009), the term 'bullshit' is now recognized as more than just a mild expletive. Bullshit is an important technical concept in management theory for understanding how to comprehend, recognize, act on, and prevent acts of communication that have no grounding in truth (Frankfurt, 2009, McCarthy et al., 2020, Spicer, 2017). Thus, bullshitting is when a human generates content not grounded in truth and then uses it for some form of social, persuasive, or evasive agenda. Bullshit is different from lying, because to lie, a liar must 'know' the truth and intentionally design statements around the truth to suit their purpose (Frankfurt, 2009). In contrast, a bullshitter is someone who has no concern for the truth and is not constrained by it. Bullshitters have the freedom to make stuff up, which can sometimes unknowingly land on the truth.

A hallucination is when an LLM generates seemingly realistic responses that are untrue, nonsensical, or unfaithful to the provided source input (Alkaissi & McFarlane, 2023). Chatbots do not 'know' the meaning of their responses, so when they generate hallucinations, they are also not lying. In contrast, the human approach to generating truthful knowledge relies on reflexivity and judgment (Lindebaum & Fleming, 2023). This does not mean that LLM generated content is always incorrect, so much as lacking the basis of a truth claim. Therefore, at best, we can consider LLM outputs as 'provisional knowledge' (Hannigan et al., 2018) in that it has no utility or impact until it is applied as part of an organizational routine or task, where legitimacy (Deephouse et al., 2017) or accountability matters (Buhmann et al., 2019). Once LLM content potentially containing a hallucination is used by a human, this application transforms it into *botshit*, which we define as LLM generated hallucinatory content that a human uncritically uses for a task.

Bullshit and botshit can come in different flavors. Pseudo-profound bullshit is statements that appear deep or meaningful at first glance but lack genuine depth and substance upon closer examination (Pennycook et al., 2015). This type of bullshit uses obscure or vague language to create an impression of wisdom or significance. For example, the statement "delivering an immersive, ultrapremium, coffee-forward experience" by the CEO of Starbucks (2017) uses grandiose-sounding but abstract phrases to create an illusion of profound meaning. Littrell et al. (2021a) suggest that bullshitting can be persuasive or evasive in nature. Persuasive bullshitting involves embellishing or stretching the truth to persuade, to impress, or fit in. It includes pseudo-profound bullshit, as people make vacuous, buzzword-riddled, empty, but appealing proclamations to sway audiences. Evasive bullshitting is a strategic circumvention of the truth. It involves making statements to avoid revealing that you do not know something or to hide something you should not have done. There is also social bullshit, the "banter, the loose talk, the unsubstantiated opinions, and the fanciful claims that lubricate and amplify our interactions with friends and family" (McCarthy et al., 2020: 256).

Understanding the nature of bullshit helps us understand the motives, impact, and responses to dealing with statements. This logic also applies to botshit, which emerging research suggests can vary depending on whether the generation of potentially hallucinatory content is an intrinsic or extrinsic infringement of its training data. Intrinsic botshit is the human application of chatbot responses that are false in that they contradict the LLM's training data (Ji et al., 2023). This type of botshit can be relatively easy for humans to identify and prevent. For example, consider a chatbot trained with accurate, up-to-date data about the micro-blogging platform X (formerly called Twitter) and its competitors. If, on December 10, 2023, a prompt asked this chatbot, "Who is the CEO of the social media platform X?" and it replied 'Elon Musk' then this response would be an intrinsic hallucination in that it contradicts the up to current date and accurate training data, which would specify 'Linda Yaccarino'. If the chatbot user uses this intrinsic hallucination response for a task, it becomes intrinsic botshit. In contrast, extrinsic botshit is when chatbot users use a response that can neither be supported or refuted by the training data (Maynez et al., 2020). For instance, if a chatbot was asked, 'What is the future business model for the social media platform X?' the non-hallucinatory response should be 'I do not know'. However, if the chatbot generated a response to this prompt, regardless of how coherent and appealing it might be, it would likely be an extrinsic hallucination, as the strategic plans for X are secret and could not be part of the LLM training data. Once this made-up response is used by someone it becomes extrinsic botshit.

The reasons that humans produce and spread bullshit and botshit are likely somewhat similar. Research suggests that people are more likely to bullshit when the social or professional expectations to have an opinion are high and when the bullshitter expects the veracity of their statements to be accepted unchallenged (Petrocelli, 2018). In other words, humans can feel that making stuff up is safer or more rewarding than saying, 'I do not know', especially if their bosses also regularly generate and spread bullshit (Ferreira et al., 2022). Similarly, Spicer (2020) explains that bullshitting in the workplace is a social practice. It can be a required 'language game' to fit in, get things done, and enhance one's standing. This behavior, in turn, signals others to engage in bullshit to attain such rewards. These conforming, persuasive, and evasive reasons for why humans generate and use bullshit also likely apply to why humans use and transform chatbot hallucinations into botshit. However, they are not behind why chatbots produce hallucinations. Chatbots should not possess intentional motives or deception agendas. Rather, a chatbot's propensity to produce hallucinations is due to the inherent nature limitations of AI and LLMs. These limitations might seem to reduce as training data and transformer algorithms improve to produce new-generation chatbots that can better handle the nuances of prompts, but AI and LLMs still only engage in memorization, not true human-like intelligence. The creative conjectures from LLMs do not involve thought-based criticism, fallible knowing, and moral thinking (Chomsky, Roberts, & Watumull, 2023; Lindebaum & Fleming, 2023).

Bullshit and botshit are likely to be more believable and impactful when they satisfy three criteria (McCarthy et al., 2020). First, it is useful, beneficial, or energizing for the audience. Second, it aligns with and flatters the audience's interests, beliefs, experiences, or attitudes. Third, it is perceived to have some credibility based on how articulate it is, the extent to which it is riddled with impressive jargon, and the perceived authority of the person or chatbot generating it. For bullshit, this last point is known as the 'Einstein Effect', as we are more likely to believe a bullshit statement if we think it is made by someone with prestigious scientist-like standing (Hoogeveen et al., 2022). Perceptions of legitimacy around the source of a statement affect attributions of plausibility of the statement's contents (Deephouse et al., 2017). Similarly, responses from a chatbot built and run by an organization with a respectable and trustworthy reputation are more likely to be believed than responses from an unknown new organization or an established organization with a track record for providing misinformation.

In sum, research on bullshit helps us to understand the motives, types, and impacts of botshit. This understanding can guide efforts to verify the chatbot response veracity. Furthermore, strategies for avoiding or lessening harm from botshit will be guided by the work context in which they are created and used. These insights motivate and guide the ideas that underlie the typology framework we present for understanding how to leverage chatbots for various work tasks while mitigating the associated botshit risks.

| | **Bullshit** | **Botshit** |
|---|---|---|
| **Defined** | Human-generated content that has no regard for the truth which a human then applies for communication and decision-making tasks (Frankfurt, 2009, McCarthy et al., 2020, Spicer, 2017).<br>For example, a human produces a report using evidence that they have made up and is untrue, and the report is presented to others. | Chatbot generated content that is not grounded in truth (i.e., hallucinations) and is then uncritically used by a human for communication and decision-making tasks.<br>For example, a human produces a report using chatbot generated content that is untrue, and the report is presented to others. |
| **Types** | *Pseudo-profound bullshit*: statements that seem deep and meaningful (Pennycook et al. 2015)<br>*Persuasive bullshit*: statements that aim to impress or persuade (Littrell et al. 2021a)<br>*Evasive bullshit*: statements that strategically circumvent the truth (Littrell et al. 2021a)<br>*Social bullshit*: statements that tease, exaggerate, joke, or troll (McCarthy et al., 2020; Spicer, 2017) | *Intrinsic botshit*: the human application of a chatbot response that contradicts the chatbot's training data (Ji et al., 2023; Sun et al., 2023)<br>*Extrinsic botshit*: the human application of a chatbot response that cannot be verified as true or false by the chatbot's training data (Ji et al., 2023; Sun et al., 2023; Maynez et al., 2020) |
| **Insights** | Humans are more likely to generate and use bullshit:<br>• The more unintelligent, dishonest, and insincere they are (Littrell et al., 2021b).<br>• The expectations for them to have an opinion are high, and they expect to get away with it (Petrocelli, 2018). | Chatbots are more likely to generate hallucinations for humans to use and transform into botshit when there are:<br>• Data collection, preprocessing and tokenization problems limit factual knowledge *alignment* between the |

| | | |
|---|---|---|
| | • If their bosses frequently spout bullshit (Ferreira et al., 2022).<br>Humans are more likely to believe and spread bullshit:<br>• If they have a low capacity for analytical thinking (Pennycook et al., 2015).<br>• If they think it is made by a scientist (Hoogeveen et al. 2022).<br>• If it is appealing, aligned with existing beliefs, and seems credible (McCarthy et al., 2020). | training data and the desired response (Sun et al., 2023).<br>• Ambiguous prompts misdirect the chatbot (White et al., 2023).<br>• Problems with the training and modeling choices of the LLM transformer (Raunak et al., 2021).<br>• Issues with fine-tuning efforts (Ramponi, 2022) based on uncertainty around ground truth (Lebovitz et al., 2021). |

Table 2: Bullshit versus botshit

### 3.1 A typology of chatbot work modes

To help reduce and avoid the epistemic risks of using chatbot-generated content for work tasks, we now explain how users should consider the modes of work they use chatbots for. To do this, we draw on risk management research that advocates the importance of assessing the impact and detectability of potential risks to deal effectively with risks (Berg, 2010; Hopkin, 2018). Thus, we suggest chatbot users consider two questions when using this technology for different work: How important is chatbot response veracity for the task? And how easy is it to verify the veracity of the chatbot response? The answer to these two questions leads to four different modes of working with responses generated by a chatbot (see Figure 1).

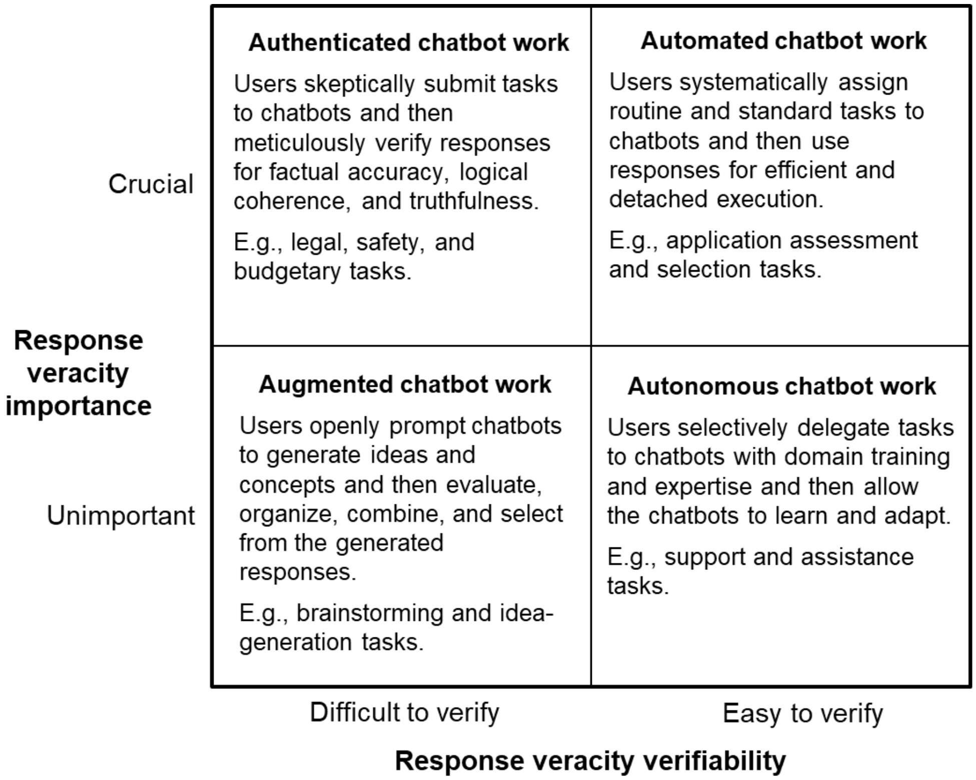
The first question, about the importance of chatbot response veracity to a task, highlights that different types of work can have different expectations or requirements for a chatbot to avoid the risk of using hallucinatory content. This precautionary approach to assessing and dealing with risks ensures that any response is suited to the severity of the risk (Oehmen et al., 2014; Carbone & Tippett, 2004). Botshit risk severity is concerned with the impact, consequence, or the amount at stake from potentially using chatbot hallucinations for different types of work.

When the risk severity of using chatbot content for work is catastrophic, chatbot response veracity is *crucial*. For instance, situations such as high-stakes investment decisions, mission critical operations activities, and situations with low or zero tolerance for failure, such as equipment maintenance, patient wellbeing, and employee safety. In this sense, risks around botshit mirror the concept of algorithmic accountability in organizations, where organizations need to manage reputational concerns, moral responsibility, and trust with stakeholders (Buhmann et al., 2019). Simply stating that an algorithm was behind a decision or operation is inadequate if a situation goes awry. Catastrophic risks associated with chatbot content mean that response veracity is crucial. Although AI designers increasingly appreciate the importance of ‘human-in-the-loop’ learning processes (Mosqueira-Rey et al., 2023), it is not sufficient to just build this into the front end of model construction and tuning (Ramponi, 2022). Response veracity refers to validation processes that are akin to fact-checking. This may increase the costs and inefficiencies of using LLMs, but risk severity also points to costs of chatbot hallucinations causing failures in high-stakes work.

When the risk severity of using chatbot content for work is low, chatbot response veracity is *unimportant*. For example, if a chatbot is used for a task that is cheap and easy to reverse, or if the effects of hallucinatory content are trivial, then the chatbot response veracity is likely to be of less importance. This might be the case if a manager sought ideas or suggestions (brainstorming or ‘botstorming’), feedback on creative work, or prompts for low-cost experiments. The ideas generated by chatbots for this type of work would be unlikely to be used as is. Instead, the chatbot generated ideas would trigger insights that help open-up a problem or solution space to understand an issue from different perspectives. Chatbots could also be safely used for idea mash-ups, where a prompt asks what can be learned from mixing two different business practices or industries. For example, the prompt “How might airport security teams learn from casinos?” was submitted to ChatGPT-4. This prompt generated a response that Casinos have long been at the forefront of using real-time monitoring and data analytic technologies for behavior analysis, facial recognition, and crowd management. While these are just ideas based on word patterns, like ideas from a human, they still should be assessed, modified, and combined as needed before being used.

The second question that users need to consider before using a chatbot is how straightforward it is to check and verify the veracity of a chatbot’s response. In risk management research, this is an issue of risk detection (Carbone & Tippett, 2004; Hopkin, 2018), i.e., what is the likelihood that inaccuracies and made-up content can be identified in a chatbot response? If hallucinatory content is easy to check and identify, then risk detection wise, this is like a runaway train that can be heard for miles (Pritchard, 2000). Thus, the verifiability of a chatbot response indicates how likely a user can identify hallucinatory content. This depends on how simple-complex, cheap-expensive, and quick-slow, it is to review, discover and prevent hallucinatory content in a chatbot response from becoming used and transformed into botshit.

Responses will be *easy* to verify if a relatively settled and accessible set of truth claims exists around the response content. For example, a prompt to a chatbot asking for the definition of a word can be easily verifiable using a relatively stable set of truth sources. These sources include reputable online dictionaries that make it relatively straightforward to verify definition-oriented responses. Moreover, online multilingual translation services such as Google Translate can check the accuracy of responses involving translation work quickly and confidently.



| Response veracity importance | Difficult to verify | Easy to verify |
|---|---|---|
| Crucial | **Authenticated chatbot work**<br>Users skeptically submit tasks to chatbots and then meticulously verify responses for factual accuracy, logical coherence, and truthfulness.<br>E.g., legal, safety, and budgetary tasks. | **Automated chatbot work**<br>Users systematically assign routine and standard tasks to chatbots and then use responses for efficient and detached execution.<br>E.g., application assessment and selection tasks. |
| Unimportant | **Augmented chatbot work**<br>Users openly prompt chatbots to generate ideas and concepts and then evaluate, organize, combine, and select from the generated responses.<br>E.g., brainstorming and idea-generation tasks. | **Autonomous chatbot work**<br>Users selectively delegate tasks to chatbots with domain training and expertise and then allow the chatbots to learn and adapt.<br>E.g., support and assistance tasks. |

**Response veracity verifiability**

Figure 1: Four modes of chatbot work

Relatively *difficult* veracity verifiability is where it is costly and time-consuming to assemble truth claims around the response content. This would include prompting chatbots to work on tasks related to competitor analysis, technology audits, regulatory change, and consumer trends. In such cases, efforts would be made to find reliable secondary data and triangulate truth claims (Denzin, 1978; Jick, 1979). Veracity verification is vital as AI companies increasingly offer products and services that work directly with organizations as clients who provide their private datasets for LLM training (Wiggers, 2023). One capability that could be particularly useful in this regard is codifying knowledge bases for *automated* or *autonomous* modes in Figure 1. For example, coding could indicate the integrity and durability (i.e., a still can-use date) of a company's data when used with chatbot technology applications and training.

There will be instances where chatbot response veracity verifiability is challenging or impossible to determine. This will be the case with work tasks where the truth is unknown, such as envisioning, scenario planning, and roadmapping, which are, to some extent, works of fiction. These are often future-oriented claims involving acts of imagination rather than accurate descriptions. Innovators, marketers, and entrepreneurs often produce these future-oriented statements that exceed known knowns but must articulate compelling and inspiring goals for their stakeholders. For example, the prompt "In ten years' time, how will generative chatbots have transformed the future of the accounting profession?" was submitted to ChatGPT-4. The response provided superficial explanations of how generative chatbots will significantly transform the accounting profession by automating client interactions, bookkeeping, compliance, and auditing services. It is difficult to immediately and confidently assess the veracity of this sort of 'predictive' response for a long-term forecast of a future industry state.

When the answers to these two questions are combined, it can help users know how to approach and use the output from four different modes of chatbot work (see Figure 1). This framework is not mutually exclusive and collectively exhaustive. Like AI automation and augmentation in general (Raisch & Krakowski, 2021), each mode of chatbot work will not always be neatly separated from each other. However, our framework still provides a grounded and useful view of modes of chatbot work to understand how to mitigate the epistemic risks associated with botshit.

When the response veracity of a task is difficult to verify and the response veracity is unimportant, then a user should use the response generated by the chatbot in an *augmented* way. By this, we mean they should use prompted responses generated by a chatbot to enhance a user's capabilities. The chatbot could be a source that helps explore or generate ideas – perhaps as part of an ideation process. The response from the chatbot needs to be actively sifted through, questioned, edited and used as a prompt for further development. It should not be used as a final input or output for a task – instead, it is more of a creative prompt. This way of working with chatbots is likely to be the case when users try to experiment and create new ideas at the start of a design or decision-making process. These ideas will likely be worked upon or transformed before being implemented. The result is that there is likely to be much more tolerance for testing and checking before they become costly and irreversible decisions.

When it is hard to verify the veracity of a chatbot response and response veracity is crucial, users likely need to engage in the *authenticated* mode of chatbot work. This requires users to configure extra guardrails with the chatbot and employ critical thinking skills, inductive reasoning, and forecasting to estimate the veracity of a statement produced by a chatbot. In the same way that users use real options analysis to estimate the opportunities and costs of continuing or abandoning a project (Lee & Shin, 2018), users can identify and quantify uncertainties and risks associated with the response from a chatbot. In these settings, users need to structure their commitment to the chatbot response and adapt the responses as (un)certainties unfold. This helps ensure they make better decisions. An example of this use of a chatbot is a user planning a high-stakes investment into a new industry about which there is a lack of up-to-date and accurate information. The importance of veracity in the decision is crucial, but the verifiability of the response is hard– simply because there is a lack of information that can be checked against. In this context, while chatbot responses can inform decision-making, they should be thoroughly reviewed and validated using critical thinking. This allows for better evaluation of potential solutions and risk assessment. Chatbots can highlight overlooked details or problems, adding depth and thoroughness to critical decisions, especially in uncertain or ambiguous situations.

If it is easy to verify a response whose veracity is relatively crucial, then this would be the *automated* mode of chatbot work. Users carefully assign simple, routine, and relatively standardized tasks to a chatbot. Because the veracity of the content generation is vital, users would need to sense-check the responses from chatbots before they are implemented. This is a form of quality control rather than a co-creation process (as takes place with the authenticating mode of chatbot work). An example of this kind of chatbot work would be the analysis and pre-approval of loan applications in a bank. In these cases, large amounts of information could be used to verify the statements made. However, the decision (particularly if it is a large loan) will likely be relatively important and high stakes. This means that while the chatbot might automatically produce a veracious statement, it still needs to be checked and approved by a trained banker. With this mode of chatbot work, technology is speeding up much of the routine work that goes into making verifiable decisions.

If it is possible to easily verify a response whose veracity is relatively unimportant, then this would be an *autonomous* mode of chatbot work. This mode involves users selectively delegating to chatbots specific tasks. These chatbots might have domain-based training to learn from and adapt to. Being able to assign these low-stakes verifiable tasks to chatbots is likely to come with a potential for increased productivity and removing routine tasks from individuals. Examples of this autonomous mode of chatbot include responding to routine customer inquiries or processing common administrative inquiries. With such examples, the chatbots would also autonomously be able to redirect complex inquiries they can't handle to a human who might work in conjunction with augmented or authenticated chatbot modes or provide an answer.

## 4.0 USING CHATBOTS WITH INTEGRITY

We are currently in an era of exploration, where different fields are experimenting with different possibilities for integrating chatbot into work. As a result, some fields will see this technology supplant jobs and tasks (Dwivedi et al., 2023), and others will see it being used as a valuable sidekick or co-pilot (Glaser & Gehman, 2023). This latter perspective corresponds with recent work in management on "conjoined agency" as the "shared capacity between humans and nonhumans to exercise intentionality" (Murray et al., 2021:555) and speaks to the generative potential of a people and AI intertwined. When using chatbots, as outlined in Figure 1, users should be aware that each mode comes with a specific epistemic risk. Chatbot users should design processes and practices to manage these risks. In this section, we look at the epistemic risks linked to each mode – *ignorance* for augmented, *miscalibration* for authenticated, *routinization* for automated, and *black boxing* for autonomous - and suggest how they can be managed. While each type of risk could appear in all forms of chatbot work, we suggest each type of risk will be most acute with a specific mode of chatbot work due to response veracity verifiability and response veracity importance. These ideas help ensure that this AI technology is used for integrity-driven value creation from a work process and content perspective (Canhoto & Clear, 2020).

With the augmented mode of chatbot work, perhaps ignorance is the most important risk. This happens when chatbot users overlook or are unaware of the technology's potentially useful and harmful outputs. This happens when users blindly rely on the technology in a limited way. They are relatively closed in their view of the value and hazards that come from chatbot work. This can mean they have less information and capacity to process it than they might otherwise have. To deal with this risk of ignorance, users should work on ways to ensure that people are prompted to incorporate outputs produced by chatbots into their decision-making. One way to do this is by building chatbots into routine decision-making practices. For instance, a chatbot could also be consulted in addition to consulting stakeholders or experts. When sending out communications, the message could be pre-checked and edited using a chatbot explicitly trained for a firm's corporate communications. Another aspect of ignorance happens when information created by a chatbot is accurate but entirely overlooked by decision-makers because it is perceived as inaccurate by its source. This can be dealt with through anonymization or even personifying information from the chatbot. Doing this could be treated as equal input to other experts or stakeholders. Through managing the risk of ignorance, managers can ensure that relevant information from chatbots is fed into decision-making processes.

The primary risk that comes with the authenticated mode is that users may miscalibrate the value of chatbot responses for their work. This miscalibration can happen in two ways. One is that users systematically view chatbot responses as having more veracity and value for work than they do and do not authenticate enough. Second, they excessively distrust the veracity and value of chatbot responses, resulting in high authentication and response rejection levels. Both forms of miscalibration can lead decision-makers to systematically under or over utilize chatbot responses in contrast to other sources of information (such as their own expert judgment). One way to deal with these calibration risks is to systematically calculate and track the veracity and value of chatbot outputs for work. The calculation could be based on past likelihood of making correct judgments or a judgment of how likely to present judgment is correct. Such a process is recommended as a best practice in forecasting (Tetlock & Gardner, 2016).

With the automated mode of chatbot work the central risk is the level of routinization employed. For example, if work tasks become over-routinized, then decision-makers could lose over-sight of the work being automated and effectively 'fall asleep at the wheel' (Dell'Acqua, 2021). This might be acceptable if the chatbot is handling standard problems that it is well-trained to deal with. However, more user oversight and less routinization will be required if chatbot work tasks become more varied and unconventional in scope. For these reasons, ensuring that the user overseeing automated chatbot work does not lose focus and 'drift' is essential. One way to mitigate this risk is by requiring this chatbot work mode to be periodically accompanied by manual work and engagement (even though it is not strictly needed). This places the user into a copiloting role requiring attention and engagement. Users remain alert and engaged and effectively monitor the output of automated chatbot work.

The chief risk with the autonomous mode of chatbot work is understanding the extent to which the chatbot is black boxed (Latour, 1987), i.e., the internal workings of the chatbot are not fully understood or accessible to the user. Users often do not care to understand how technology does its magic-like work because the business model and algorithms behind the technology can be secret, opaque, inaccessible, and fixed. This means users are likely to face epistemic risks because they are unaware of the limits and potential of the chatbot technology they use. To guard against this black boxing risk, users could be required to learn how their chatbot's work. This could involve producing a version of Table 1 in this paper specific to a company's chatbot training data and task scope. However, be careful and note that to some extent, the black boxing of a chatbot could be beneficial as it inhibits users from gaming or sabotaging the technology for personal gain or perverse agendas.

## 5.0 LEARN TO RELY ON ME

When the portable calculator was introduced into U.S. schools in the mid-1970s, there were initial concerns about how it would impact the teaching profession and students' learning development (Banks, 2011). However, these fears soon started to disappear. Eventually, the calculator became widely used in classrooms to complement many, but not all, aspects of a student's math education. Chatbot technologies are currently undergoing a similar development and adoption journey. They provide similar magical-like assistance but for content creation and analysis, and they are increasingly being used for work and educational tasks. Where they differ from calculators is in the type of knowledge they generate, and the potential epistemic risks associated with botshit. Thus, there will always be a risk that users could contaminate their work with untruths and inaccuracies produced by chatbots. Hence, we now present some guardrails (i.e., rules, guidelines, or limitations for chatbot use) for how the technology, organizations, and users can mitigate these risks and enhance the truthiness of chatbot use for work.

**Technology-oriented guardrails** - As outlined in Table 1, these focus on the technical aspects and capabilities of a chatbot and its LLM. They help ensure that the mechanics and scope of an LLM are appropriate for the mode of chatbot work it is being used for. The different modes would require different fact-checking modules within the model and chatbot tool to verify the accuracy of information before responding. For example, the automated and autonomous modes would require the most specialized, fixed, and strongest technology-oriented guardrails to ensure sufficient trust to use chatbots for these work modes. Technology-oriented guardrails would also include cross-reference routines that train the LLM data using trusted sources, flagging (or correcting false or misleading responses), and assessing the credibility of data sources. Specific applications and modes of chatbot work (i.e., authenticated and augmented) would also benefit more from real-time updates of the LLMs training data to ensure that responses are accurate and up-to-date.

**Organization-oriented guardrails** – These are the guidelines and policies organizations develop to mitigate the risk of botshit. Like a code of conduct, organization-oriented guardrails help an organization understand and demand appropriate and acceptable use of chatbots to ensure veracity, integrity, and responsible use of chatbot generated content. These guardrails involve employee training on the capabilities and limitations of chatbots while considering the kinds of tasks for which they use chatbots and associated botshit risks (i.e., the chatbot work modes in Figure 1). The guidelines would outline how chatbot technology would be implemented and evaluated to ensure trustworthy use and prescribe regular safety checks to review and audit the veracity of chatbot-generated content. They would also provide prompt engineering training (White et al., 2023) to formulate effective prompts for the response veracity verifiability, and response veracity importance of each mode of chatbot work.

Furthermore, as per failure management research, when botshit is produced and consumed, there would be rules to promote transparency and disclosure to properly investigate, learn from, and prevent future failures (Cannon & Edmondson, 2005). Recently, two types of transparency were found to be effective for building confidence in this technology: “explanations of how the algorithm works and reflection of AI reliability” (Glikson & Woolley, 2020: 648). Organization-oriented guardrails could be operationalized using Simons’ (1994) ‘four levers of control system’ where an organization’s (i) beliefs systems ensure a vision and values for mitigating botshit, (ii) the boundary systems specify and enforce limits about how and when to use chatbots to mitigate risks, (iii) diagnostic systems determine and support measurable limits of botshit use in the workplace; and (iv) and interactive systems promote feedback, learning and innovation in pursuit of botshit free use of chatbots in the workplace.

**User-oriented guardrails** – These concern the abilities and practices that human users would exhibit to help mitigate the risks of botshit in the workplace. As per research on understanding and dealing with workplace bullshit (McCarthy et al., 2020; Spicer, 2017), different levels of critical thinking and fact-checking would be expected for each of the four modes of chatbot work in our typology. The authenticated mode would benefit most from users having a critical mindset that questions, cross-references, and validates chatbot generated content to minimize botshit risks. Meanwhile, the augmented mode is more likely to reward users with a less critical and more open mindset, i.e., a mentality that leverages, adapts, and builds on the content rather than using it as is. Finally, in line with risk management research (Edmondson, 2002), to mitigate the hazards of botshit, users should have the courage and responsibility to speak up and question the veracity of the chatbot responses.

## 6.0 FINAL THOUGHTS

The computer scientist David Mimno (2023) recently compared chatbots with historical soothsayers. He pointed to the classic work of Ibn Khaldûn in describing the soothsayer as "often speak[ing] the truth and agrees with reality. Often, however, what he says are falsehoods, because he supplements his deficiency with something foreign to, different from, and incompatible with, his perceptive essence. Thus, truth and falsehood are jumbled together in him" (Khaldûn, 2020: 80). Our paper explains that when this jumble of truth and falsehood is used for work tasks, it can become botshit. For chatbots to be used reliably, it is important to recognize that their responses can best be thought of as provisional knowledge (Hannigan, et al., 2018). In contrast to validated knowledge derived from legitimate sources (Mulkay, 1979), provisional knowledge is an open innovation related concept where organizations grapple with situations rife with ambiguity and learn to rely on knowledge sources with debated veracity (Deephouse et al., 2017). To master chatbot-generated provisional knowledge and mitigate the risks of possible botshit, we provide a framework of four modes of chatbot work and related advice to avoid blindly using chatbot predictions.

**Acknowledgments**

We thank CK Kaligotla, Yisheng Li, and Vern Glaser for their valuable input that helped us frame and develop this paper.

## References

Alkaissi, H., & McFarlane, S. I. (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*. https://doi.org/10.7759/cureus.35179

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction machines: the simple economics of artificial intelligence*. Harvard Business Press.

Babic, C. (2019). A Theory of Epistemic Risks. *Philosophy of Science*, 86(3): 522-550.

Banks, S. (2011). A historical analysis of attitudes toward the use of calculators in junior high and high school math classrooms in the United States since 1975. *Master of Education Research Thesis* http://digitalcommons.cedarville.edu/education_theses/31

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

Belot, H. (2023). KPMG lodges complaint after AI-generated material was used to implicate them in non-existent scandals. *The Guardian*. November 3, 2023. https://www.theguardian.com/business/2023/nov/03/kpmg-ai-complaint-non-existent-scandal-ai-case-studies-google-bard

Benioff, M. (2023) https://twitter.com/Benioff/status/1614372552025178114?lang=kn January 14, 2023

Berg, H. P. (2010). Risk management: procedures, methods and experiences. *Reliability: Theory & Applications*, 5(2 (17)), 79-95.

Brown., L. (2023). Thousands of authors including Atwood, Egan and Picoult sign AI open letter. The Bookseller. July 19, 2023. https://www.thebookseller.com/news/thousands-of-authors-including-atwood-egan-and-picoult-sign-ai-open-letter

Buhmann, A., Paßmann, J., & Fieseler, C. (2019). Managing Algorithmic Accountability: Balancing Reputational Concerns, Engagement Strategies, and the Potential of Rational Discourse. *Journal of Business Ethics*. https://doi.org/10.1007/s10551-019-04226-4

Canhoto, A. I., & Clear, F. (2020). Artificial intelligence and machine learning as business tools: A framework for diagnosing value destruction potential. *Business Horizons*, *63*(2), 183-193.

Cannon, M. D., & Edmondson, A. C. (2005). Failing to learn and learning to fail (intelligently): How great organizations put failure to work to innovate and improve. *Long range planning*, *38*(3), 299-319.

Carbone, T. A., & Tippett, D. D. (2004). Project risk management using the project risk FMEA. Engineering management journal, 16(4), 28-35.

Chandrasekar, P. (2023). Announcing OverflowAI - stack overflow. https://stackoverflow.blog/2023/07/27/announcing-overflowai/. Accessed on 06-11-2023.

Chen, L., Zaharia, M., & Zou, J. (2023). *How is ChatGPT's behavior changing over time?* (arXiv:2307.09009). arXiv. https://doi.org/10.48550/arXiv.2307.09009

Chesbrough, H. W. (2003). *Open innovation: The new imperative for creating and profiting from technology*. Harvard Business Press.

Chomsky, N., Roberts, I., & Watumull, J. (2023). Noam Chomsky: The False Promise of ChatGPT. *The New York Times*, *8*.

Chui, M., Roberts, R., & Yee, L. (2022). Generative AI is here: How tools like ChatGPT could change your business. *Quantum Black AI by McKinsey*.

Dell'Acqua, F. (2021). Falling asleep at the wheel: Human/AI collaboration in a field experiment on HR recruiters. Retrieved from: https://static1.squarespace.com/static/604b23e38c22a96e9c78879e/t/61a85f09c5599734019f714c/1638424329219/Fabrizio+DellAcqua+-+Falling+Asleep+at+the+Wheel+-+Dec+2.pdf

Deephouse, D.L., Bundy, J., Tost, L.P., Suchman, M.C. (2017). Organizational legitimacy: six key questions. In: Greenwood, R., Oliver, C., Lawrence, T., Meyer, R. (Eds.), *The SAGE Handbook of Organizational Institutionalism*, 2nd ed. Sage publications, Thousand Oaks, CA

Denzin, N. K. (1978) *The Research Act*, 2d ed. New York: McGraw-Hill.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Edmondson, A. C. (2002). *Managing the risk of learning: Psychological safety in work teams* (pp. 255-275). Cambridge, MA: Division of Research, Harvard Business School.

Ferreira, C., Hannah, D., McCarthy, I., Pitt, L. & Lord Ferguson, S., (2022). This place is full of it: Towards an organizational bullshit perception scale. *Psychological Reports*, *125*(1), pp.448-463.

Frankfurt, H. G. (2005). *On Bullshit*. Princeton University Press.

Glaser, V.L. & Gehman, J. (2023). Chatty Actors: Generative AI and the Reassembly of Agency in Qualitative Research. *Journal of Management Inquiry*. In-press

Glikson, E. & Woolley, A.W. (2020). Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals, 14(2), pp.627-660.

Guardian (2023) https://www.theguardian.com/commentisfree/2023/apr/06/ai-chatgpt-guardian-technology-risks-fake-article

Hannigan, T. R., Haans, R. F. J., Vakili, K., Tchalian, H., Glaser, V. L., Wang, M. S., Kaplan, S., & Jennings, P. D. (2019). Topic Modeling in Management Research: Rendering New Theory from Textual Data. *Academy of Management Annals*, 13(2), 586–632

Hannigan, T. R., Seidel, V. P., & Yakis Douglas, B. (2018). Product innovation rumors as forms of open innovation. *Research Policy*, 47(5), 953–964

Hoogeveen, S., Haaf, J.M., Bulbulia, J.A., Ross, R.M., McKay, R., Altay, S., Bendixen, T., Berniūnas, R., Cheshin, A., Gentili, C. and Georgescu, R. (2022). The Einstein effect provides global evidence for scientific source credibility effects and the influence of religiosity. *Nature Human Behaviour*, 6(4), pp.523-535.

Hopkin, P. (2018). Fundamentals of risk management: understanding, evaluating and implementing effective risk management. Kogan Page Publishers.

Jiao, W., Wang, W., Huang, J. T., Wang, X., & Tu, Z. (2023). Is ChatGPT a good translator? A preliminary study. arXiv preprint arXiv:2301.08745.

Jarrahi, M.H., (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision-making. *Business Horizons*, 61(4), pp.577-586.

Jarrahi, M.H., Askay, D., Eshraghi, A. and Smith, P. (2023). Artificial intelligence and knowledge management: A partnership between human and AI. *Business Horizons,* 66(1), pp.87-99.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12), 248:1-248:38. https://doi.org/10.1145/3571730

Jick, T. D. (1979). Mixing Qualitative and Quantitative Methods: Triangulation in Action. *Administrative Science Quarterly*, 24(4), 602–611.

Khaldûn, I. (2020). *The muqaddimah: an introduction to history-abridged edition*. Princeton University Press.

Kietzmann, J., Lee, L.W., McCarthy, I.P. and Kietzmann, T.C. (2020). Deepfakes: Trick or treat? *Business Horizons*, 63(2), pp.135-146.

Kietzmann, J. and Pitt, L.F. (2020). Artificial intelligence and machine learning: What managers need to know. *Business Horizons*, 63(2), pp.131-133.

Klein, E. (2023). The Ezra Klein Show. https://www.nytimes.com/2023/01/06/podcasts/transcript-ezra-klein-interviews-gary-marcus.html, January 6, 2023

Kozyrkov, C. (2022). What Is 'Ground Truth' in AI? (A Warning.). *Medium*. August 19, 2022. https://towardsdatascience.com/in-ai-the-objectiveis-subjective-4614795d179b

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Harvard University Press.

Lebovitz, S., Levina, N., & Lifshitz-Assaf, H. (2021). Is AI ground truth really true? The dangers of training and evaluating ai tools based on experts' know-what. *MIS Quarterly*, 45(3), 1501–1525.

Lee, I. & Shin, Y.J. (2018). Fintech: Ecosystem, business models, investment decisions, and challenges. *Business Horizons*, 61(1), pp.35-46.

Lin, S., Hilton, J. and Evans, O., (2021). Truthfulqa: Measuring how models mimic human falsehoods. *preprint arXiv*:2109.07958.

Lindebaum, D., & Fleming, P. (2023). ChatGPT Undermines Human Reflexivity, Scientific Responsibility and Responsible Management Research. *British Journal of Management*. https://doi.org/10.1111/1467-8551.12781

Littrell, S., Risko, E. F., & Fugelsang, J. A. (2021). 'You can't bullshit a bullshitter'(or can you?): Bullshitting frequency predicts receptivity to various types of misleading information. *British journal of social psychology 60*(4), 1484-1505.

Littrell, S., Risko, E.F. and Fugelsang, J.A. (2021). The bullshitting frequency scale: Development and psychometric properties. *British Journal of Social Psychology*, 60(1), pp.248-270.

Mauran, C. (2023). OpenAI is being sued for training ChatGPT with 'stolen' personal data. Mashable. June 29, 2023. https://mashable.com/article/openai-chatgpt-class-action-lawsuit

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

McCarthy, I. P., Hannah, D., Pitt, L. F., & McCarthy, J. M. (2020). Confronting indifference toward truth: Dealing with workplace bullshit. *Business Horizons*, *63*(3), 253-263.

Mimno, D. (2023) https://twitter.com/dmimno/status/1681286888731336704?s=12&t=N29uXkkupWUW3qvFoSR46Q, July 18, 2023

Mollick, E. (2023) https://twitter.com/emollick/status/1729358803425001702, November 27, 2023.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, 56(4), 3005–3054. https://doi.org/10.1007/s10462-022-10246-w

Mukherjee, A., & Chang, H. H. (2023). Managing the Creative Frontier of Generative AI: The Novelty-Usefulness Tradeoff. *California Management Review Insights*. https://cmr.berkeley.edu/2023/07/managing-the-creative-frontier-of-generative-ai-the-novelty-usefulness-tradeoff/

Mulkay, M. (1979). Knowledge and utility: Implications for the sociology of knowledge. *Social Studies of Science*, *9*(1), 63-80.

Munn, L., Magee, L., & Arora, V. (2023). Truth Machines: Synthesizing Veracity in AI Language Models. *arXiv preprint arXiv:2301.12066*.

Murray, A., Rhymer, J. E. N., & Sirmon, D. G. (2021). Humans and technology: Forms of conjoined agency in organizations. *Academy of Management Review*, *46*(3), 552-571.

New York Times (2023) https://www.nytimes.com/2023/06/22/nyregion/lawyers-chatgpt-schwartz-loduca.html

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Available at SSRN 4375283*.

O'Brian, M. (2023). ChatGPT-maker OpenAI signs deal with AP to license news stories. *Associated Press.* Retrieved from https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a

Oehmen, J., Olechowski, A., Kenley, C. R., & Ben-Daya, M. (2014). Analysis of the effect of risk management practices on the performance of new product development programs. *Technovation*, *34*(8), 441-453.

OpenAI (2023) GPT-4 System Card https://cdn.openai.com/papers/gpt-4-system-card.pdf

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S., Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, & R. Lowe. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730-27744.

Paschen, U., Pitt, C., & Kietzmann, J. (2020). Artificial intelligence: Building blocks and an innovation typology. *Business Horizons*, *63*(2), 147-155.

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2015). On the reception and detection of pseudo-profound bullshit. *Judgment and Decision-making*, *10*(6), 549-563.

Petrocelli, J. V. (2018). Antecedents of bullshitting. *Journal of Experimental Social Psychology*, *76*, 249-258.

Pritchard, C. L. (2000). *Risk Management: Concepts and Guidance*. CRC Press.

Przegalinska, A., Ciechanowski, L., Stroz, A., Gloor, P. and Mazurek, G. (2019). In bot we trust: A new methodology of chatbot performance measures. *Business Horizons*, 62(6), pp.785-797.

Reuters (2023) https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. *Academy of management review*, *46*(1), 192-210.

Ramponi, M. (2022). ‘How ChatGPT actually works’, Retrieved from https://www.assemblyai.com/blog/how-chatgpt-actually-works/ .

Raunak, V., Menezes, A., & Junczys-Dowmunt, M. (2021). The curious case of hallucinations in neural machine translation. *arXiv preprint arXiv:2104.06683*.

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?. *Journal of Applied Learning and Teaching*, *6*(1).

Shum, H. Y., He, X. D., & Li, D. (2018). From Eliza to XiaoIce: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, *19*, 10-26.

Simons, R. (1994) How new top managers use control systems as levers of strategic renewal. *Strategic Management Journal*, 15, 169–189

Spicer, A. (2017). *Business Bullshit.* Routledge.

Spicer, A. (2020). Playing the bullshit game: How empty and misleading communication takes over organizations. *Organization Theory*, 1(2), p.2631787720929704.

Sun, W., Shi, Z., Gao, S., Ren, P., de Rijke, M., & Ren, Z. (2023). Contrastive learning reduces hallucination in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 37, No. 11, pp. 13618-13626).

Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The art and science of prediction*. Random House.

Vincent J. (2022) AI-Generated Answers Temporarily Banned on Coding Q&A Site Stack Overflow. *The Verge*. December 5, 2022. https://www.theverge.com/2022/12/5/23493932/chatgpt-ai-generated-answers-temporarily-banned-stack-overflow-llms-dangers

Wallace, R.S. (2009). *The anatomy of ALICE* (pp. 181-210). Springer Netherlands.

Weil, E. (2023). ‘You Are Not a Parrot And a chatbot is not a human. And a linguist named Emily M. Bender is very worried what will happen when we forget this’, New York Magazine, 1 March.

Weizenbaum, J. (1996) ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*. 9, no. 1 (1966): 36-45.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT* (arXiv:2302.11382). arXiv. https://doi.org/10.48550/arXiv.2302.11382

Wiggers, K. (2023). OpenAI wants to work with organizations to build new AI training data sets. *Tech Crunch*. November 9, 2023. https://techcrunch.com/2023/11/09/openai-wants-to-work-with-organizations-to-build-new-ai-training-data-sets

Xiao, Y., & Wang, W. Y. (2021). On hallucination and predictive uncertainty in conditional language generation. *arXiv preprint arXiv:2103.15025*.

Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., & Hashimoto, T. B. (2023). Benchmarking large language models for news summarization. arXiv preprint arXiv:2301.13848.

Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z. and Du, Y. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.