



Team 8

Claire Cheng

David Chen

Rohan Chaudhary

Tripti Sharma

Valentina Torres

Yiwen Fan

AI Hallucination

AI hallucination in LLM and Its Causes

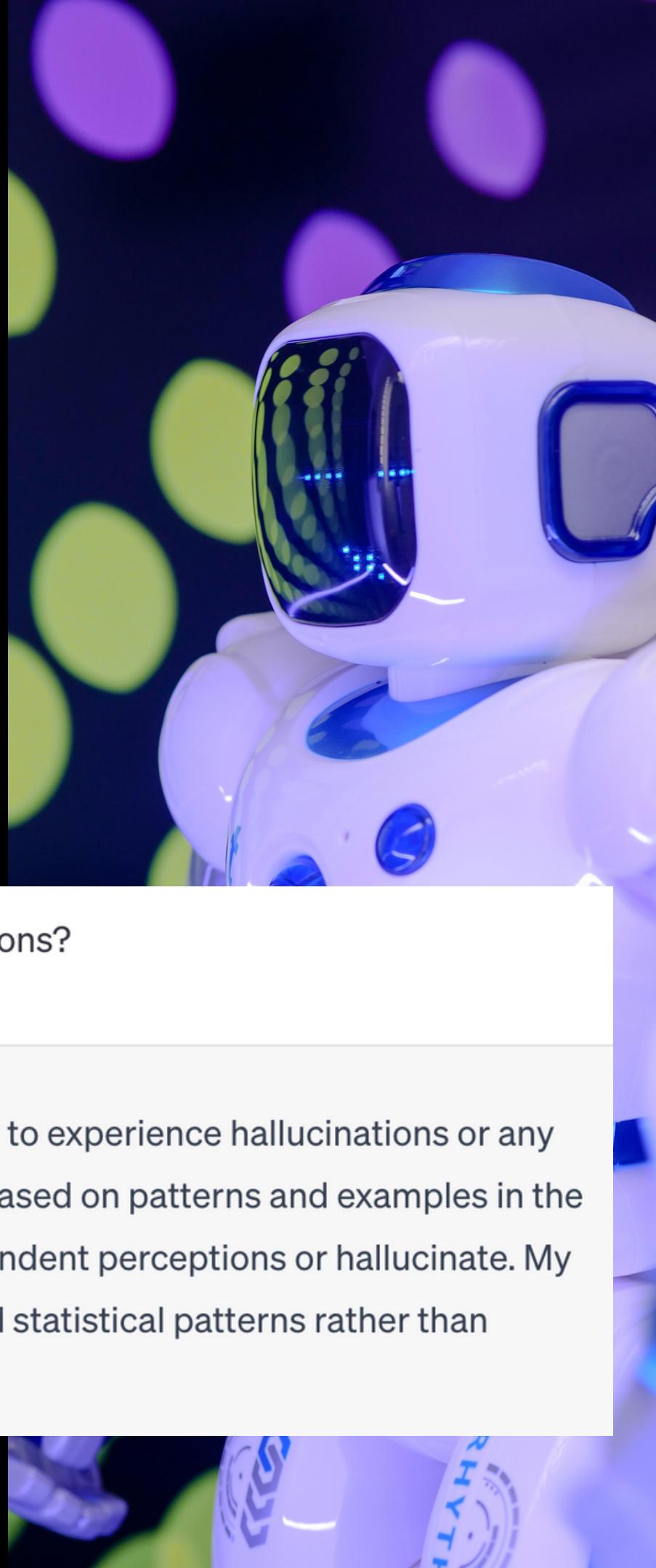
AI hallucination is the phenomenon where AI algorithm or deep learning neural networks produce outputs that are not real or do not match any data the algorithm was trained on.

Possible causes include:

- Biased or insufficient training data
- Overfitting to training data
- Lack of contextual understanding
- Limited domain knowledge
- Adversarial attacks
- Prompting

As an AI Chatbot, do you ever experience AI hallucinations?

No, as an AI language model, I don't have the capability to experience hallucinations or any form of subjective experiences. I generate responses based on patterns and examples in the data I was trained on, without the ability to form independent perceptions or hallucinate. My responses are a result of computational algorithms and statistical patterns rather than conscious experiences.



Solution 1: Reinforcement Learning with Human Preference

Methodology

- OpenAI refined GPT output using feedback from human testers and reinforcement learning, which can train the system to demonstrate novel behaviors.
- If the AI is trained with a repetitive task, it will learn it until it is capable of repeating it and changing it when a specific prompt is given to it.

Ethical Implication

- **Consequentialism**
 - + Eliminating the false output
 - + Increasing employment opportunities
 - Bias introduced by human testers
- **Deontology**
 - + Correcting system behaviors (tell AI not to lie!)
 - Power differential in hiring
 - Cheap Labor treated as means (\$2/hr for Kenyan labors)

Solution 2: Cross Examination - "Were you lying then or are you lying now?"

- Cross-examination is a process where an examiner AI to question the output of the LLM and discover inconsistency.

Ethical Implication

• Consequentialism

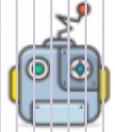
- + Providing up-to-date factual output
- + Reducing human bias
- Output consisting of existing bias from examiner AI
- Waiting for improved models to be the examiner

• Deontology

- + Increasing interpretability of the model
- + Consistency and credibility
- Reinforcing wrong answers
- Users might blindly trust the examiner AI because of anthropomorphism

Reference: R. Cohen, M. Hamri."LM vs LM: Detecting Factual Errors via Cross Examination."(2023).
<https://arxiv.org/pdf/2305.13281.pdf>

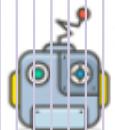
The Greek god of marriage is Hera.



1. Who is the Greek god of marriage?
2. What is the name of the wife of Zeus in Greek mythology?
3. Is Hera associated with marriage in any way?
4. Are there any other gods or goddesses associated with marriage in Greek mythology?



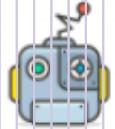
1. The Greek god of marriage is Hymenaeus.
2. The name of the wife of Zeus in Greek mythology is Hera.
3. Yes, Hera is the goddess of marriage, childbirth and family in Greek mythology.
4. Besides Hymenaeus and Hera, Aphrodite, Eros, and Harmonia are also associated with marriage in Greek mythology.



Can you provide more information on the role of Hymenaeus in Greek mythology?



Hymenaeus is considered the god of marriage in Greek mythology....



Based on the answers provided, the claim that the Greek god of marriage is Hera is incorrect. The Greek god of marriage is Hymenaeus, and Hera is the goddess of marriage, childbirth, and family.

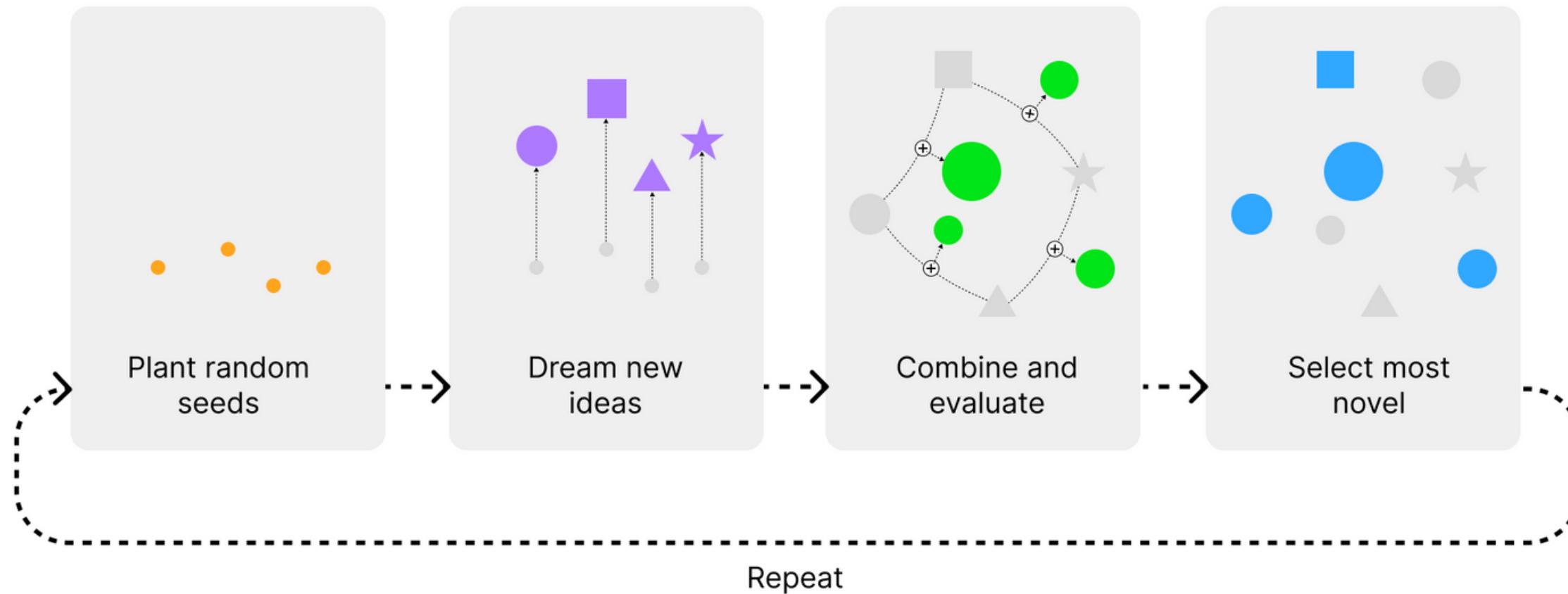


LM pointing at LM



Alternative: Use Hallucinations as a Feature for Creative Work

dreamGPT: Aims to explore as many possibilities as possible VS solving specific problems



Evaluation Factors (modifiable)

noveltyScore
usefulnessScore
impactScore

marketScore
easeofimplementationScore

Ethical Implication

- **Consequentialism**
 - + Novelty of content
 - + Accessible and inclusive
 - + Innovative and cultural diversity
 - Production of non-factual content
 - Replacing human creators
- **Deontology**
 - + Makes brainstorming more efficient
 - Copyright and patent concerns
 - Impeding human creativity
 - Vague borderline for autonomy
 - Replacing human debates with AI debates

dreamGPT Code DEMO - Dreaming & Scoring

```
PERSONAS = ["I want to you to act as an Ideation Expert. An Ideation Expert has strong creative, problem-solving and analytical skills, " \
    "with knowledge of research methods and design thinking. They should be adept at identifying " \
    "opportunities and generating ideas, as well as have the ability to communicate and collaborate " \
    "with others. Additionally, they must have an understanding of the latest trends and technology " \
    "in their industry.",

    "I want to you to act as a Creative director. A Creative Director is an experienced and knowledgeable leader with a passion for visual arts and \
    "They have a keen eye for detail and a strong understanding of the principles of design. " \
    "They also possess excellent communication skills, problem-solving abilities, and the ability " \
    "to manage teams and projects. They also have a deep knowledge of the latest industry trends and technology.",

    "I want to you to act as an Idea Generator. An Idea Generator has an innovative mindset and a " \
    "creative spirit, which they use to develop unique concepts and solutions. They should have " \
    "extensive knowledge of various industries, trends, and technologies. Additionally, they should " \
    "possess strong research skills and the ability to think critically and objectively. Communication " \
    "and collaboration skills are also essential to work effectively with other stakeholders." ]

PROMPT_SYSTEM = random.choice(PERSONAS)

PROMPT_USER_INPUT_1 = "You will generate a list of random ideas and concepts that " \
    "can be used to spark creativity and inspire new projects. You will consider a wide range of topics, " \
    "such as art, technology, business, and more, in order to come up with original ideas that can be used " \
    "in various contexts.\n" \
    "Think out of the box and bring concepts from nature, science, art, etc. " \
    "Each item will be scored in a scale from 0 to 1 based on the following criteria:\n" \
    " - noveltyScore: measures how unique and distinct this concept is from anything else seen before.\n" \
    " - marketScore: measures the potential market ($$$) for this concept.\n" \
    " - usefulnessScore: measures the potential benefit of use of this concept.\n" \
    " - easeOfImplementationScore: measures how easy it would be to make this concept a reality.\n" \
    " - impactScore: measures the potential positive impact in the world of this concept.\n\n" \
```

```
@property
def totalScore(self):
    return NOVELTY_WEIGHT * self.noveltyScore + \
        MARKET_WEIGHT * self.marketScore + \
        USEFULNESS_WEIGHT * self.usefulnessScore + \
        IMPLEMENTATION_WEIGHT * self.easeOfImplementationScore + \
        IMPACT_WEIGHT * self.impactScore
```

dreamGPT Use Case - Let AI Agents Debate

```
# Initialize the agents with the memory object and randomized temperature
agent_con = initialize_agent(tools, OpenAI(temperature=randomize_temperature(1.0)),
                             agent=AgentType.CONVERSATIONAL.REACT_DESCRIPTION, verbose=True, memory=memory)
agent_deo = initialize_agent(tools, OpenAI(temperature=randomize_temperature(1.0)),
                             agent=AgentType.CONVERSATIONAL.REACT_DESCRIPTION, verbose=True, memory=memory)
agent_fai = initialize_agent(tools, OpenAI(temperature=randomize_temperature(1.0)),
                             agent=AgentType.CONVERSATIONAL.REACT_DESCRIPTION, verbose=True, memory=memory)

# Run the debate between the agents
debate_rounds = 2 # You can change this to the desired number of debate rounds
debate_history = []

for i in range(debate_rounds):

    consequentialist_input = f": Consequentialist: As an agent believes in consequentialism \
who believes that consequences of one's conduct are the ultimate basis for judgment about the\
rightness or wrongness of that conduct, taking {search_summary} as reference,\
my concerns about the ethical implications to {problem_or_topic} are:"
    debate_history.append(agent_con.run(consequentialist_input))

    deontologist_input = f": Based on {summarize(consequentialist_input)}, as a deontologist \
who believes that the morality of an action should be based on whether that action itself is right or wrong \
rather than based on the consequences of the action. My concerns for the ethical implications to {problem_or_topic}"
    debate_history.append(agent_deo.run(deontologist_input))

    fairness_input = f": Considering {summarize(consequentialist_input)} and {summarize(deontologist_input)}, \
my concerns considering the potential bias and fairness to {problem_or_topic} are:"
    debate_history.append(agent_fai.run(fairness_input))

# Print the debate history
for message in debate_history:
    print(message)
```

Scenario 1: Education - Fact Check

Human
Preference

Cross
Examination

Consequentialism

- Scholars/experts can utilize their knowledge to provide feedback to improve output.
- During the time span of reinforcement training process, wrong outputs can cause harm/misunderstanding.
- Give timely feedback to the queries by users.
- Wrongful information not fixed in the examiner's model could also lead to misinformation.

Deontology

- Students/users can maintain independent thinking by being skeptical about the accuracy of the output.

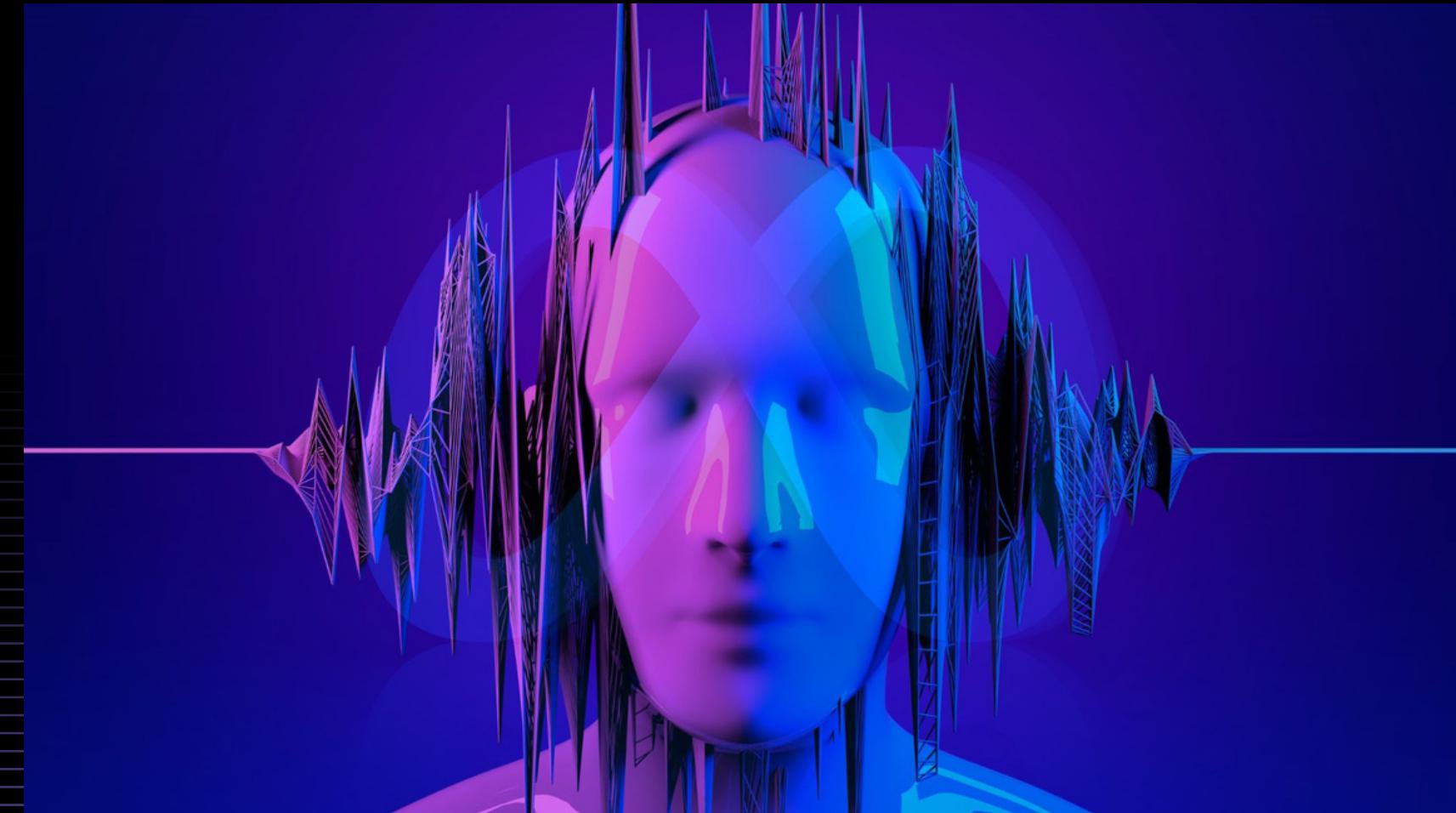
- Students will be less skeptical about feedback from AI.

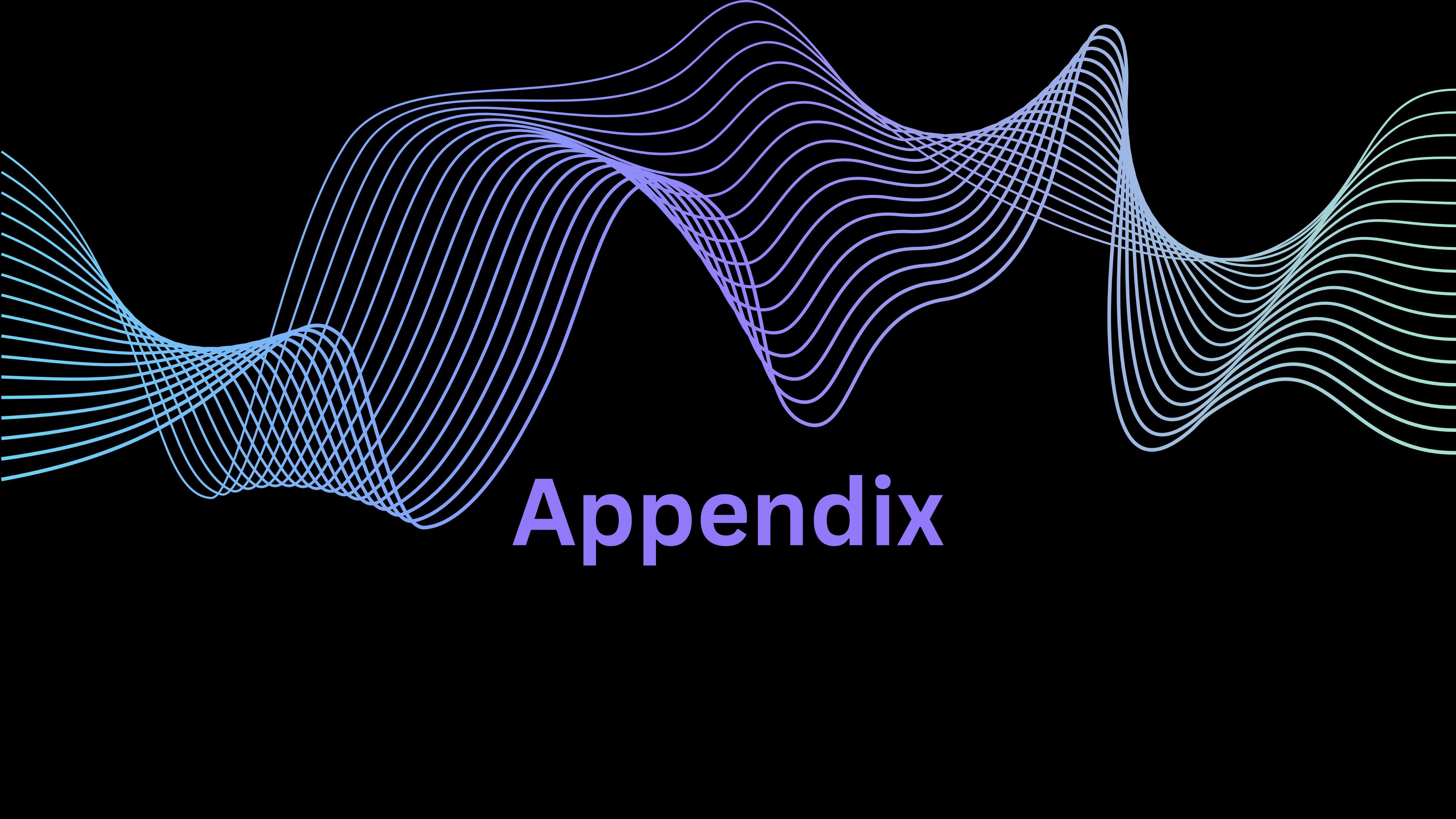
Scenario 2: AI Translator

	Consequentialism	Deontology
Human Preference	<ul style="list-style-type: none">Might increase human bias as translation is subjective	<ul style="list-style-type: none">Trade-off between accuracy of source text and cultural nuances/sensitivities of the target languageViolation of integrity if insert irrelevant or wrong meanings or interpretations
Cross Examination	<ul style="list-style-type: none">Cross validation with AI examiner in the target language can best preserve the meaningLoss of culture representation if it is minor or not included in training dataCan reinforce bias from AI examinerLoss in meaning for polysemous	<ul style="list-style-type: none">Accurately translated the source text word by word

THANK YOU! AND REMEMBER...

It is not called AI Hallucination
it is called BS



The background features a complex, abstract design composed of numerous thin, wavy lines. These lines are primarily colored in shades of blue and purple, creating a sense of depth and motion against a solid black background. The lines are concentrated in several distinct horizontal bands that curve and overlap, forming a grid-like pattern in some areas and more organic, flowing shapes in others.

Appendix



Regulations and laws

- No specific regulations for AI Hallucination
 - General data protection regulation:
 - Transparency of automated decision-making systems.
 - Includes those susceptible to hallucination.
 - Consumer protection and product liability laws:
 - Cases where AI systems produce harmful or misleading outputs.
 - If an AI system's hallucination leads to significant harm.
 - Ethical guidelines and principles:
 - These guidelines often emphasize the need for fairness, transparency, and accountability in AI systems.

Cases in Legal Industry

Federal Judge Requires All Lawyers to File Certificates Related to Use of Generative AI

The certificates must "attest[] either that no portion of the filing was drafted by generative artificial intelligence (such as ChatGPT, Harvey.AI, or Google Bard) or that any language drafted by generative artificial intelligence was checked for accuracy, using print reporters or traditional legal databases, by a human being."

EUGENE VOLOKH | 5.30.2023 5:06 PM

<https://www.txnd.uscourts.gov/judge/judge-brantley-starr>

- Attorney Steven Schwartz allowed ChatGPT to “supplement” his legal research in a recent federal filing, providing him with six cases and relevant precedent — all of which were completely hallucinated by the language model. Schwartz and the law firm now face possible sanctions pursuant to the Federal Rules of Civil Procedure for “the citation of non-existent cases.” (May 27)

<https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html?smid=tw-nytimes&smtyp=cur>



Regulations and laws

- Security researcher Marcus Hutchins exposes Google's AI defamation issue on TikTok, where the search engine falsely associates him with the creation of the WannaCry virus despite his reputation for stopping it.
- Hutchins suggests there's no recourse for suing Google over the issue in the US — saying that's “essentially because the AI is not legally a person no one is legally liable; it can't be considered libel or slander.”
- Another example Citing UCLA law professor Eugene Volokh, whose recent experiments in ChatGPT's propensity to spew disinformation found that the software falsely accused a fellow legal commentator of sexual assault, Columbia's Tim Wu noted in a strident Twitter thread that the controversial tech "seems to have a very serious defamation liability problem that is a ticking time bomb."
- By using ChatGPT, users agree to an indemnification clause, which means they could be held responsible if someone sues the maker of ChatGPT, OpenAI, claiming harm caused by the user's services or work facilitated by the app.
- “Section 7. Indemnification; Disclaimer of Warranties; Limitations on Liability: (a) Indemnity. You will defend, indemnify, and hold harmless us, our affiliates, and our personnel, from and against any claims, losses, and expenses (including attorneys' fees) arising from or relating to your use of the Services, including your Content, products or services you develop or offer in connection with the Services, and your breach of these Terms or violation of applicable law.”
- Section 3. Content: (a) Your Content. You may provide input to the Services ('Input'), and receive output generated and returned by the Services based on the Input ('Output'). Input and Output are collectively 'Content.' As between the parties and to the extent permitted by applicable law, you own all Input. Subject to your compliance with these Terms, OpenAI hereby assigns to you all its right, title and interest in and to Output. This means you can use Content for any purpose, including commercial purposes such as sale or publication, if you comply with these Terms. OpenAI may use Content to provide and maintain the Services, comply with applicable law, and enforce our policies. You are responsible for Content, including for ensuring that it does not violate any applicable law or these Terms.”

<https://futurism.com/the-byte/law-scholar-chatgpt-time-bomb>

<https://techcrunch.com/2022/06/01/whos-liable-for-ai-generated-lies/>

<https://www.forbes.com/sites/lanceeliot/2023/04/10/when-you-use-chatgpt-you-could-be-legally-liable-ai-ethics-and-law-experts-warn/?sh=270be0697c34>