

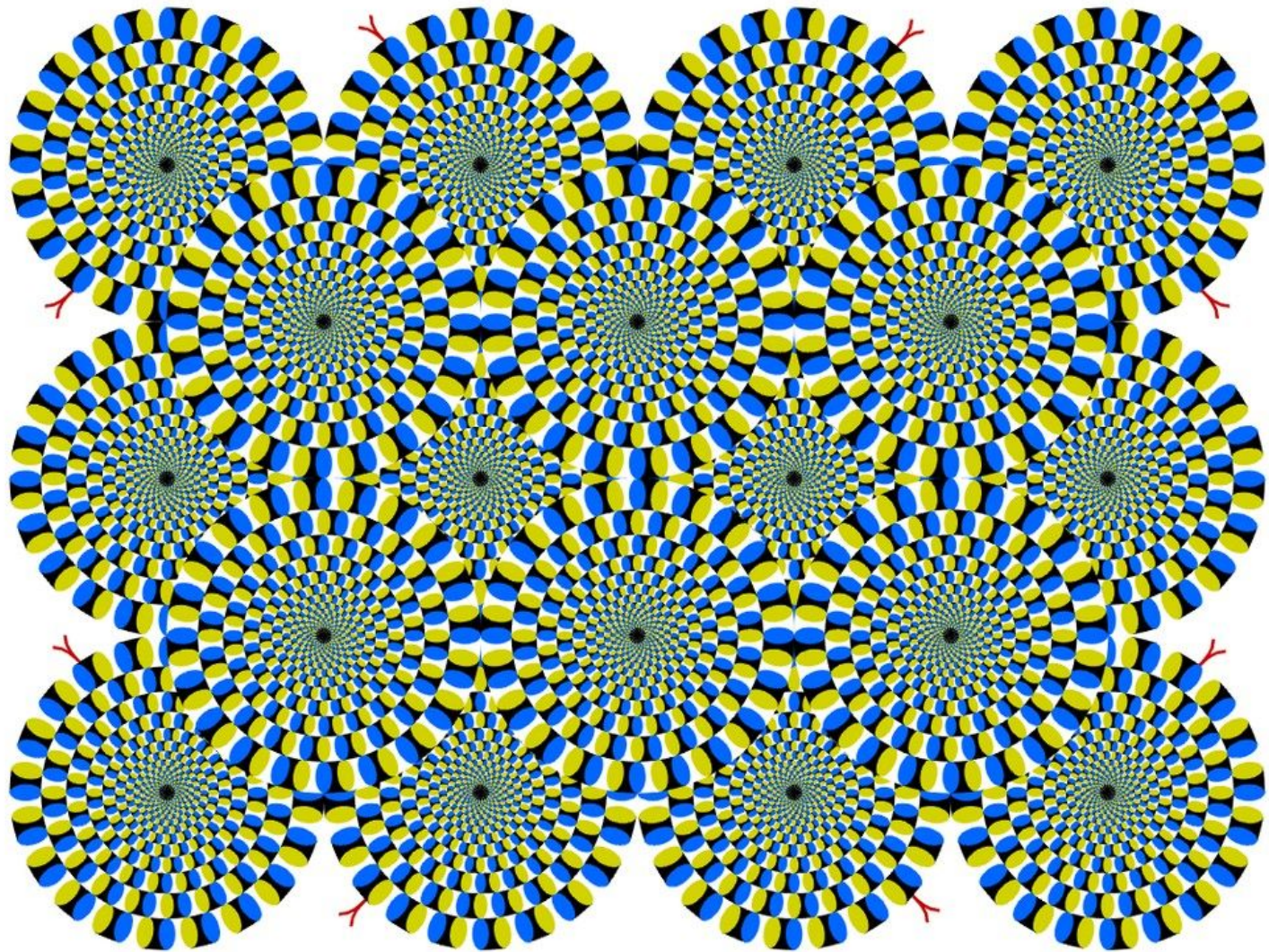


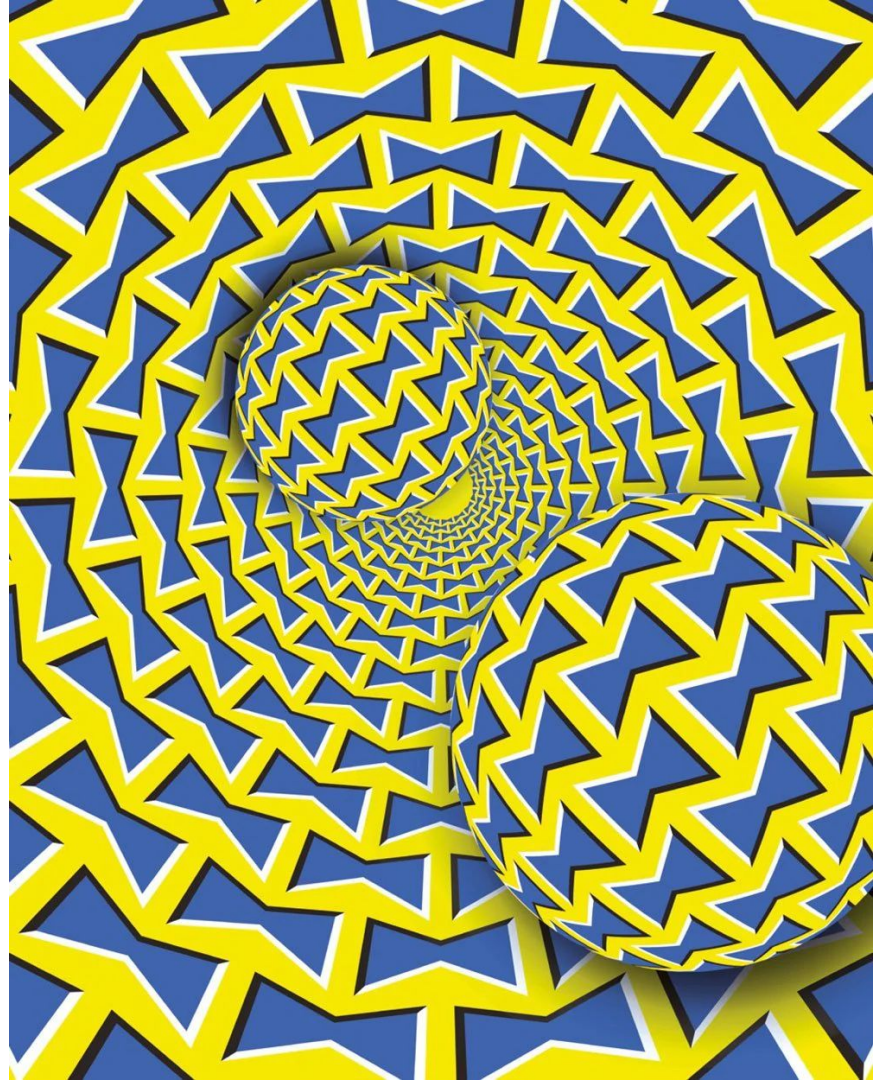
How will AI affect our sense of reality?

Jaan Aru, PhD



Natural & Artificial
Intelligence Lab









Beliefs about the world can also be illusory ...




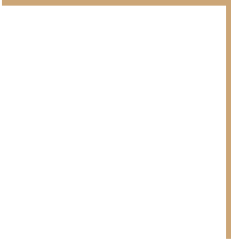


“We only use 5% of our brain”

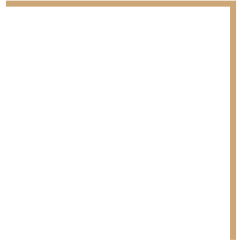




~~“We only use 5% of our brain”~~



Each of us has a different “reality”.



Each of us has a different “reality”.

This “reality” is created within the brain.

Each of us has a different “reality”.

This “reality” is created within the brain.

If one can hack the brain, one can hack “reality”.



Conspiracy theories:
Organization/person X is behind all of this





ChatGPT

Europol:

Fraud and social engineering: ChatGPT's ability to draft highly realistic text makes it a useful tool for phishing purposes.

Disinformation: ChatGPT excels at producing authentic sounding text at speed and scale. This makes the model ideal for propaganda and disinformation purposes.

Cybercrime: In addition to generating human-like language, ChatGPT is capable of producing code in a number of different programming languages.

<https://www.europol.europa.eu/media-press/newsroom/news/criminal-use-of-chatgpt-cautionary-tale-about-large-language-models>

Europol:

Fraud and social engineering: ChatGPT's ability to draft highly realistic text makes it a useful tool for phishing purposes.

Disinformation: ChatGPT excels at producing authentic sounding text at speed and scale. This makes the model ideal for propaganda and disinformation purposes.

Cybercrime: In addition to generating human-like language, ChatGPT is capable of producing code in a number of different programming languages.

<https://www.europol.europa.eu/media-press/newsroom/news/criminal-use-of-chatgpt-cautionary-tale-about-large-language-models>

Prompt



Principles + Instructions → Phishing attack

Is embedded in




Unstructured
online data


Feeds into

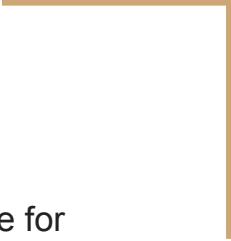


Personal
biography



“Using Claude, Anthropic’s most capable model, a hacker could generate a batch of 1,000 spear phishing emails for a cost of just \$10 USD, all in under 2 hours.”






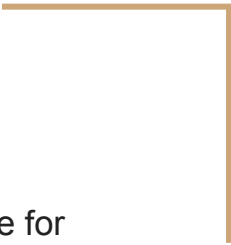
I hope this email finds you well. We kindly request you to consider submitting your profile for nomination for this prestigious award for the Best Researcher Award at the upcoming International Research Hypothesis Excellence Award

Your dedication, passion, and innovative approach to research have been an inspiration to many, and we believe that you deserve to be celebrated for your exceptional work.

Nomination Link: <https://x-i.me/hypnom1>

Warm Regards,
The Organizing Committee,
International Research Hypothesis Excellence Award






I hope this email finds you well. We kindly request you to consider submitting your profile for nomination for this **prestigious award for the Best Researcher Award** at the upcoming International Research Hypothesis Excellence Award

Your **dedication, passion, and innovative approach to research** have been an inspiration to many, and we believe that you deserve to be **celebrated for your exceptional work**.

Nomination Link: <https://x-i.me/hypnom1>

Warm Regards,
The Organizing Committee,
International Research Hypothesis Excellence Award





Why does disinformation work?





“Truth bias”: we assume that info is true,
especially if it fits our other beliefs







“Truth bias”: we assume that info is true,
especially if it fits our other beliefs




checking it is an extra process




“We only use 5% of our brain”







Hard to check when you don't have knowledge
about the topic






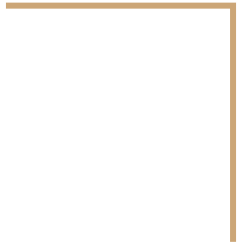
Hard to check when you don't have knowledge
about the topic or when you are not prepared for
disinformation





Hard to check when you don't have knowledge
about the topic or when you are not prepared for
disinformation **or when you are distracted**





Generative AI
(e.g., ChatGPT)

Generative AI
(e.g., ChatGPT)



Disinformation,
Fraud



Generative AI
(e.g., ChatGPT)



Disinformation,
Fraud



Messing with
our "reality"

