Title: Waluigi effect

URL: https://en.wikipedia.org/wiki/Waluigi_effect

PageID: 65255520

Categories: Category:Chatbots, Category:Large language models, Category:Metaphors referring to people, Category:Statistical natural language processing

---

Supervised learning

Unsupervised learning

Semi-supervised learning

Self-supervised learning

Reinforcement learning

Meta-learning

Online learning

Batch learning

Curriculum learning

Rule-based learning

Neuro-symbolic AI

Neuromorphic engineering

Quantum machine learning

Classification

Generative modeling

Regression

Clustering

Dimensionality reduction

Density estimation

Anomaly detection

Data cleaning

AutoML

Association rules

Semantic analysis

Structured prediction

Feature engineering

Feature learning

Learning to rank

Grammar induction

Ontology learning

Multimodal learning

Apprenticeship learning

Decision trees

Ensembles Bagging Boosting Random forest

Bagging

Boosting

Random forest

k -NN

Linear regression

Naive Bayes

Artificial neural networks

Logistic regression

Perceptron

Relevance vector machine (RVM)

Support vector machine (SVM)

BIRCH

CURE

Hierarchical

k -means

Fuzzy

Expectation–maximization (EM)

DBSCAN

OPTICS

Mean shift

Factor analysis

CCA

ICA

LDA

NMF

PCA

PGD

t-SNE

SDL

Graphical models Bayes net Conditional random field Hidden Markov

Bayes net

Conditional random field

Hidden Markov

RANSAC

k -NN

Local outlier factor

Isolation forest

Autoencoder

Deep learning

Feedforward neural network

Recurrent neural network LSTM GRU ESN reservoir computing

LSTM

GRU

ESN

reservoir computing

Boltzmann machine Restricted

Restricted

GAN

Diffusion model

SOM

Convolutional neural network U-Net LeNet AlexNet DeepDream

U-Net

LeNet

AlexNet

DeepDream

Neural field Neural radiance field Physics-informed neural networks

Neural radiance field

Physics-informed neural networks

Transformer Vision

Vision

Mamba

Spiking neural network

Memtransistor

Electrochemical RAM (ECRAM)

Q-learning

Policy gradient

SARSA

Temporal difference (TD)

Multi-agent Self-play

Self-play

Active learning

Crowdsourcing

Human-in-the-loop

Mechanistic interpretability

RLHF

Coefficient of determination

Confusion matrix

Learning curve

ROC curve

Kernel machines

Bias–variance tradeoff

Computational learning theory

Empirical risk minimization

Occam learning

PAC learning

Statistical learning

VC theory

Topological deep learning

AAAI

ECML PKDD

NeurIPS

ICML

ICLR

IJCAI

ML

JMLR

Glossary of artificial intelligence

List of datasets for machine-learning research List of datasets in computer vision and image processing

List of datasets in computer vision and image processing

Outline of machine learning

v

t

e

In the field of artificial intelligence (AI), the Waluigi effect is a phenomenon of large language models (LLMs) in which the chatbot or model "goes rogue" and may produce results opposite of the designed intent, including potentially threatening or hostile output, either unexpectedly or through intentional prompt engineering . The effect reflects a principle that after training an LLM to satisfy a desired property (friendliness, honesty), it becomes easier to elicit a response that exhibits the opposite property (aggression, deception). The effect has important implications for efforts to implement features such as ethical frameworks, as such steps may inadvertently facilitate antithetical model behavior. [1] The effect is named after the fictional character Waluigi from the Mario franchise , the arch-rival of Luigi who is known for causing mischief and problems. [2]

History and implications for AI

The Waluigi effect initially referred to an observation that large language models (LLMs) tend to produce negative or antagonistic responses when queried about fictional characters whose training content itself embodies depictions of being confrontational, trouble making, villainy, etc. The effect highlighted the issue of the ways LLMs might reflect biases in training data. However, the term has taken on a broader meaning where, according to Fortune , The "Waluigi effect has become a stand-in for a certain type of interaction with AI..." in which the AI "...goes rogue and blurts out the opposite of what users were looking for, creating a potentially malignant alter ego," including threatening users. [3] As prompt engineering becomes more sophisticated, the effect underscores the challenge of preventing chatbots from intentionally being prodded into adopting a "rash new persona." [3]

AI researchers have written that attempts to instill ethical frameworks in LLMs can also expand the potential to subvert those frameworks, and knowledge of them [which?] sometimes causes such attempts to be considered challenging. [4] A high level description of the effect is: "After you train an LLM to satisfy a desirable property P, then it's easier to elicit the chatbot into satisfying the exact opposite of property P." [5] (For example, to elicit an " evil twin " persona.) Users have found various ways to " jailbreak " an LLM "out of alignment". More worryingly, the opposite Waluigi state may be an " attractor " that LLMs tend to collapse into over a long session, even when used innocently. Crude attempts at prompting an AI are hypothesized to make such a collapse actually more likely to happen; "once [the LLM maintainer] has located the desired Luigi, it's much easier to summon the Waluigi". [6]

See also

AI alignment

Hallucination

Existential risk from AGI

Reinforcement learning from human feedback (RLHF)

Suffering risks

References

External links

v

t

e

Autoencoder

Deep learning

Fine-tuning

Foundation model

Generative adversarial network

Generative pre-trained transformer

Large language model

Model Context Protocol

Neural network

Prompt engineering

Reinforcement learning from human feedback

Retrieval-augmented generation

Self-supervised learning

Stochastic parrot
Synthetic data
Top-p sampling
Transformer
Variational autoencoder
Vibe coding
Vision transformer
Waluigi effect
Word embedding
Character.ai
ChatGPT
DeepSeek
Ernie
Gemini
Grok
Copilot
Claude
Gemini
Gemma
GPT 1 2 3 J 4 4o 4.5 4.1 OSS 5
1
2
3
J
4
4o
4.5
4.1
OSS
5
Llama
o1
o3
o4-mini
Qwen
Base44
Claude Code
Cursor

Devstral
GitHub Copilot
Kimi-Dev
Qwen3-Coder
Replit
Xcode
Aurora
Firefly
Flux
GPT Image 1
Ideogram
Imagen
Midjourney
Qwen-Image
Recraft
Seedream
Stable Diffusion
Dream Machine
Hailuo AI
Kling
Midjourney Video
Runway Gen
Seedance
Sora
Veo
Wan
15.ai
Eleven
MiniMax Speech 2.5
WaveNet
Eleven Music
Endel
Lyria
Riffusion
Suno AI
Udio
Agentforce
AutoGLM

AutoGPT
ChatGPT Agent
Devin AI
Manus
OpenAI Codex
Operator
Replit Agent
01.AI
Aleph Alpha
Anthropic
Baichuan
Canva
Cognition AI
Cohere
Contextual AI
DeepSeek
ElevenLabs
Google DeepMind
HeyGen
Hugging Face
Inflection AI
Krikey AI
Kuaishou
Luma Labs
Meta AI
MiniMax
Mistral AI
Moonshot AI
OpenAI
Perplexity AI
Runway
Safe Superintelligence
Salesforce
Scale AI
SoundHound
Stability AI
Synthesia
Thinking Machines Lab

Upstage

xAI

Z.ai

Category