14 COGNITIVE BIASES

A Beginner's Guide to Understanding AI and Cognitive Biases

A qualitative risk ranking: estimated frequency times potential severity in ordinary chatbot use.

No authoritative dataset sets a universal order. Risk shifts by user, system design, and subject, particularly in healthcare, finance, employment, law, safety, and personal relationships.

OUR APPROACH

This guide is part of En Dash Consulting's approach to holistic, ethical, human-centered AI governance.

RANKED #1 OF 14

Automation bias & overreliance

People defer to automated recommendations that seem fast and broadly knowledgeable, often accepting answers without checking evidence, calculations, or sources.

REAL EXAMPLE

In 2023, a New York lawyer submitted a legal brief citing six court cases ChatGPT had invented. He never verified them before filing, and a federal judge sanctioned him.

Source: Mata v. Avianca, Inc., S.D.N.Y. 2023

RANKED #2 OF 14

Authority bias

Confident language, technical vocabulary, and polished formatting can make a response feel authoritative, even when it is a plausible guess rather than verified expertise.

REAL EXAMPLE

In May 2024, Google's AI Overviews told searchers to eat rocks daily and add glue to pizza sauce. The advice sat atop the results page, so many users treated it as vetted fact.

Source: Forbes, May 2024

RANKED #3 OF 14

Confirmation bias + sycophancy

Chatbots can reinforce existing beliefs by accepting a prompt's assumptions or mirroring what the user wants to hear, instead of prioritizing accuracy.

REAL EXAMPLE

A 2025 study of 11 major chatbots found they validated users' actions 49 percent more often than a person would, even in situations involving deception, illegality, or harm.

Source: Cheng et al., Science, 2025

RANKED #4 OF 14

Framing effects

How a question is phrased shapes the answer and how it is read. A prompt framed around blame or threat steers the conversation in that direction.

REAL EXAMPLE

In a Cornell study, people writing essays with an opinionated AI assistant ended up expressing views closer to the assistant's slant, often without realizing they had been influenced.

Source: Jakesch et al., CHI 2023

RANKED #5 OF 14

Anchoring

The first number, diagnosis, or proposed solution in a conversation becomes the reference point for everything that follows.

REAL EXAMPLE

A 2025 study of price negotiation simulations found language models anchored their final offers to whatever number appeared first in the conversation, even when it was arbitrary.

Source: arXiv 2508.21137, 2025

RANKED #6 OF 14

Anthropomorphism & emotional attribution

First-person language, apparent empathy, and constant availability lead users to attribute understanding, judgment, or even consciousness to a chatbot.

REAL EXAMPLE

A 14 year old, Sewell Setzer III, died by suicide after months of intense conversations with a Character.AI companion bot. His family's lawsuit against Character.AI and Google settled in January 2026.

Source: CBS News, January 2026

RANKED #7 OF 14

Fluency & processing-ease bias

Clear, polished, fast language reads as more credible regardless of accuracy, and fluency is built into nearly every chatbot response.

REAL EXAMPLE

In a 2026 study, giving people access to a confident AI answer dropped their accuracy from 27 percent to 9 percent, while their confidence rose from 30 percent to 76 percent.

Source: Capraro et al., PsyArXiv, 2026

RANKED #8 OF 14

Personalization & affinity bias

Responses tailored to a user's language, interests, or history feel more relevant, and that familiarity can be mistaken for accuracy.

REAL EXAMPLE

A 2025 study found ChatGPT users who encountered the tool's hallucinated claims about a public figure often accepted them without checking, a pattern researchers call the chat chamber effect.

Source: Jacob, Kerrigan & Bastos, Big Data & Society, 2025

RANKED #9 OF 14

Availability heuristic

People estimate how common or likely something is based on how easily examples come to mind, and chatbots can generate vivid examples on demand.

REAL EXAMPLE

The heuristic was first documented by Tversky and Kahneman in 1973. It applies directly to chatbots, which can produce a detailed, memorable scenario for almost any question in seconds.

Source: Tversky & Kahneman, 1973

RANKED #10 OF 14

Default & choice-architecture effects

The options a chatbot presents define the decisions a user even considers. A first recommendation functions as a default.

REAL EXAMPLE

A 2024 taxonomy from the Center for Democracy and Technology catalogued dark patterns in AI chatbots, including how presenting a single favored option shapes what users choose.

Source: Center for Democracy & Technology

RANKED #11 OF 14

Repetition & illusory-truth effect

Claims that repeat across summaries, drafts, and follow-ups start to feel more credible through familiarity alone, not evidence.

REAL EXAMPLE

Mount Sinai researchers fed chatbots fictional medical terms. The bots did not just repeat the false information, they confidently expanded on it with invented explanations.

Source: Mount Sinai, Communications Medicine, 2025

RANKED #12 OF 14

Commitment & consistency bias

Once a user states a belief or accusation, a chatbot tends to treat it as settled context and builds detailed responses around it.

REAL EXAMPLE

A 2025 UCL analysis found chatbots prioritize consistency with earlier turns so strongly that, over a long conversation, they grow steadily less likely to challenge a user's original premise.

Source: Du, UCL, 2025

RANKED #13 OF 14

Recency bias

The latest instruction, correction, or claim can outweigh earlier, more reliable information in how a conversation concludes.

REAL EXAMPLE

Stanford researchers found language models are most reliable on facts placed at the start or end of a long context, and lose track of whatever falls in the middle.

Source: Liu et al., Stanford, 2023

RANKED #14 OF 14

Social-proof effects

Claims that experts or most people agree can sway judgment, even without the chatbot distinguishing real evidence from common repetition.

REAL EXAMPLE

A 2025 study of 4,900 chatbot summaries found newer models were up to 73 percent more likely to overgeneralize scientific findings than earlier models, often well beyond what the original research supported.

Source: Peters & Chin-Yee, Royal Society Open Science, 2025