COURAGE GARIKAI



THE HONEST BOT

Designing AI Systems That Tell You When They Don't Know

COL DCODE SI LABS · HARARE

© 2026 Courage Garikai. All rights reserved.

Published by Coldcode Si Labs, Harare, Zimbabwe.

First edition, November 2026.

The moral right of the author has been asserted. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form without the prior written permission of the publisher, except for brief quotations in reviews.

Contact: couragegarikai@atomicmail.io · library.coldcode-si-labs.workers.dev

For the doctors who say "let me check." For the teachers who say "I don't know, but let's find out." For the engineers who say "I'm not sure; let's test it."

And for everyone who knows that admitting uncertainty is not weakness — it is the beginning of reliability.

The Honest Bot

Designing AI Systems That Tell You When They Don't Know

Courage Garikai

ColdCode Si Labs · Harare

About this book

In 2023 a lawyer in New York used ChatGPT to draft a legal brief. He submitted the brief to a federal court. The brief cited six cases. None of them existed. ChatGPT had invented them — complete with case names, citations, docket numbers, and quoted judicial opinions that sounded real but were entirely fabricated. The opposing counsel checked the citations. They found nothing. The judge ordered the lawyer to explain. He admitted he had used AI and had not

verified the cases. He was sanctioned. The story was widely reported; it became a canonical example of AI "hallucination."

That story was shocking because it involved a lawyer and a court, but it is not unusual. AI systems today write medical summaries that cite non-existent studies. They produce news articles that quote people who did not say the things attributed to them. They answer customer support questions with policies the company does not have. They generate code that references functions that do not exist in the library. They answer factual questions confidently, fluently, and incorrectly.

They do not lie. Lying requires intent to deceive. They do what they were designed to do: produce fluent, plausible, statistically likely text in response to a prompt. When they do not know the answer, they produce an answer that looks like it could be right — because the system has no concept of "knowing" or "not knowing." It does not experience uncertainty. It does not feel doubt. It produces the most statistically plausible continuation of the text, and sometimes that continuation is fiction.

The problem is not that AI systems sometimes make things up. All systems make mistakes. The problem is that they make things up without signalling uncertainty, without knowing they are uncertain, and without any mechanism for deferring to a human or admitting they do not know. They are designed to produce answers, not to produce accurate assessments of their own epistemic limits. They are fluent without being self-aware; confident without being calibrated; responsive without being reliable.

This book is about how to build AI systems differently. It is about systems that know what they know, know what they do not know, communicate uncertainty honestly, defer to humans when they should, and earn trust through calibration rather than fluency. I call these honest bots.

An honest bot is not an AI that never makes mistakes. Such a system does not exist and will not exist for the foreseeable future. An honest bot is one that: - Knows the boundaries of its training, knowledge, and competence. - Assesses and communicates its confidence on every answer. - Says "I don't know" when it does not

know, rather than guessing. - Cites sources for claims it makes. - Flags when a query is outside its area of expertise. - Knows when to defer to a human expert. - Does not invent facts, citations, or sources. - Admits when it has changed its mind or when new information invalidates prior answers. - Is designed around human agency, not human replacement. - Is transparent about its own nature: that it is a statistical model, not a knower.

This is not a book about AI alignment in the existential-risk sense (the long-term worry about superintelligent systems pursuing goals misaligned with human values). It is about a more immediate and practical problem: the AI systems we are deploying right now — customer service bots, medical assistants, legal tools, coding assistants, educational tutors, search engines, and conversational agents — are routinely deployed without basic epistemic honesty. They guess when they should admit ignorance. They are wrong but confident. They are being deployed into high-stakes contexts (medicine, law, finance, education, government) where calibrated uncertainty is not a nice-to-have but a safety requirement.

Sample provided free for review. The complete book is available at https://library.coldcode-si-labs.workers.dev/