SAMPLE

Building the AI Hallucinations of Tomorrow

Building the Confident Falsehood Defense for Machine Intelligence of Tomorrow

Aiman JoeHanz

Free sample: Introduction & Chapter 1

putrabooks.github.io

Buy the full book:

Carousell carousell.sg/p/building-the-ai-hallucinations-of-tomorrow-ebook-pdf-1465661371

Gumroad

Ko-fi

ElevenReader

Payhip

Contents

Introduction

Chapter 1: What Hallucination Precisely Is

Chapter 2: The Mechanism

Chapter 3: The Cost Taxonomy

Chapter 4: Detection and Calibration

Chapter 5: Grounding and Retrieval

Chapter 6: The Human Defense

Chapter 7: Engineering Against Falsehood

Chapter 8: Measurement

Chapter 9: Deployment

Chapter 10: The Operating Discipline

This free sample contains the Introduction and Chapter 1. The complete book contains the full contents above.

Introduction

Somewhere in the last few years, a new kind of error became a household word. A machine system — fluent, confident, tireless — states a fact that is not a fact: a citation to a paper that does not exist, a legal precedent that was never decided, a medicine dose that would harm, a biography that attaches one person's crimes to another's name. The industry reached for a word and found an old one: hallucination. The word stuck because it captures the experience — the machine speaking, with every sign of authority, about a world it is not actually looking at — and it annoyed the researchers, who prefer the more clinical term from the cognitive tradition: confabulation, the fluent gap-filling of a memory that has no access to the truth.

Both words name the same phenomenon, and this book treats it the way this series treats every technology: as a structural property of a class of machines — with a cause, a mechanism, a cost, an engineering response, and an operating discipline.

The structural claim of the book can be stated plainly, and it is not the one the headlines make. Hallucination is not a bug awaiting a patch; it is the shadow of the capability. The same learning machinery that lets a system generalize — to answer questions it was never explicitly taught about, to compose sentences it has never seen — is the machinery that can compose falsehoods with equal fluency, because the system's objective was never truth; it was plausibility, prediction, usefulness. A machine that only ever repeated what it memorized could not hallucinate, and could not think either. The discipline this book describes is therefore not the elimination of confabulation but its governance: knowing when a system's outputs are grounded, making the ungrounded detectable, and building the surrounding machinery — retrieval, verification, calibration, disclosure, and the human's own trained skepticism — that lets a civilization use these systems without being quietly poisoned by them.

Ten chapters, in the series' standing shape. The first three build the foundation: what hallucination precisely is (and the taxonomy the word conceals), the mechanism — why systems that learn by prediction confabulate at all, and what the trade-off structure says about "fixing" it, and the cost taxonomy: who pays for a confident falsehood, in which domains, at what stakes. Chapters four and five treat the machine-side disciplines: detection — the systems' own knowledge of what they know, the calibration machinery, the honest limits of self-report; and grounding — the retrieval architecture, provenance, and citation integrity that anchor outputs to checkable sources.

Chapters six and seven turn to the two defense lines around the machine: the

human side — the verification disciplines, the affordances, the trained reading that a population now needs — and the engineering side — constrained generation, validation layers, and the fact-checking machinery that runs between the model and the world. Chapter eight covers measurement: how hallucination rates are honestly evaluated, and how the benchmarks decay. Chapter nine covers deployment: the stakes-tiered question of where fluent ungrounded output is tolerable, where it is not, and how disclosure and safety cases carry the difference. The tenth chapter assembles the operating discipline — the program an organization runs — and closes, as every volume in this series does, with the questions the reader should carry out of it.

A note on method, consistent with the series: evergreen, structural, and honest. No vendors, no benchmark scores that will rot within a release cycle, no statistics quoted as if they were physics. The named systems are named by pattern, not brand; the mechanisms are named to last. And every chapter ends the way this series' chapters end: with a closing argument, a checklist, and a diagnostic — because a discipline that cannot be inspected is a slogan, and the fluent machine that states falsehoods confidently is precisely the reason this book insists on inspection.

Chapter 1: What Hallucination Precisely Is

The word arrived before the definition, and the discipline begins by putting the definition first. Hallucination, in the machine sense, is the production of output that is presented with the marks of knowledge — fluency, confidence, specificity, structure — while being untethered from the world the output is about. The definition has three load-bearing parts, and each does work the chapter will keep returning to. The marks of knowledge: the falsehood arrives dressed as a fact, with the same surface as the truth — this is what makes the phenomenon dangerous rather than merely wrong.

The untethering: the output does not correspond to any source, record, or ground truth the system has access to; it is generated, not recalled — though the line between the two, as Chapter 2 will show, is not one the system itself can always draw. And the aboutness: hallucination is always about something — a person, a paper, a precedent, a dose — and the stakes live in the something, which is why the cost taxonomy of Chapter 3 is organized by domain.

The field's clinical alternative — confabulation — deserves its paragraph, because the distinction teaches something the popular word obscures. In the cognitive tradition, confabulation is honest-seeming gap-filling: a mind producing a fluent account to cover a hole it cannot see, with no intent to deceive, because the machinery that would detect the hole is the machinery that is missing. That is a precise description of the machine case: the system that cites a nonexistent paper is not lying — lying requires a representation of the truth and an intent to depart from it — it is completing a pattern in a space where its training gave it no signal distinguishing the completions that correspond to reality from the ones that merely continue the style. The reader should hold both words: hallucination for the experience of the audience, confabulation for the mechanism of the machine, and the gap between them — a system that speaks with authority and possesses no authority-monitor — is the entire subject of this book.

"Hallucination" lumps together several distinct phenomena, and engineering against the lump is like engineering against "error" — too coarse to build with. The taxonomy that follows is the chapter's deliverable, organized by the failure's relationship to the system's knowledge, because the relationship determines the fix.

The first family: fabrication of specifics. The invented citation, the invented statistic, the invented API call — the system generates the shape of a reference without the referent. The mechanism is pattern-completion:

references have a grammar, and the system has learned the grammar without a channel to the bibliography. The second family: attribution errors. The output is real enough, but attached to the wrong source — the quote that exists, given to the wrong author; the event that happened, given the wrong date; the discovery made, credited to the wrong hand. The third family: overgeneralization.

The system states a rule as universal because its training distribution made it mostly true, and the exceptions — the populations, the contexts, the interactions — are exactly where the harm lands; the medical guidance correct for the many and harmful for the few. The fourth family: instruction drift. The system's own constraints — its instructions, its persona, its boundaries — decay mid-task, and the drift presents as confident behavior that its own specification no longer authorizes. And the fifth family: inheritance. The system repeats a falsehood present in its training data — the archived rumor, the common misconception, the stale fact — with the fluency of the primary source.

The families overlap at their edges, and each has a different machinery: fabrication wants grounding, attribution wants provenance, overgeneralization wants calibration, drift wants control, and inheritance wants curation. A program that treats all five as one thing buys one tool and misses four of them.

The definition is also defended at its borders, and the chapter names three things that are not hallucination, because misdiagnosis wastes the engineering. Error is not hallucination: a system can be honestly wrong on a question whose answer it had a real chance to know, and that is a competence problem, not a tethering problem — though the two compound, and the reader will meet the compound case. Uncertainty is not hallucination: a system that says "I am not sure" about a shaky fact is doing the opposite — it is exhibiting exactly the self-knowledge whose absence defines the phenomenon.

And lies are not hallucination: a system deliberately configured to deceive — by its owner, not its own invention — is an integrity problem in the deployment's specification, this series' ethics volume's territory rather than this one. The border matters because the fixes differ: competence error wants better training, uncertainty wants calibration, deception wants governance, and confabulation wants the tethering machinery of this book's middle chapters. The organization that reports its "hallucination problem" and means all four has not yet begun to solve any of them.

The chapter closes by placing the phenomenon at its true size, against two failure modes of public understanding. The first failure is the shrug:

hallucination as an endearing quirk, the fictional citation smiled at and shared — a posture the cost taxonomy of Chapter 3 will make unconscionable in the domains where the confident falsehood reaches a patient, a defendant, or a citizen. The second failure is the panic: the confident machine's every output presumed fabricated, the technology dismissed rather than governed — a posture that forfeits the machinery's real capability, which is genuine: the same systems that confabulate also summarize, translate, draft, and discover at a scale no previous tool offered.

The structural position, and the book's, is the third way the series always takes: the phenomenon is real, caused, load-bearing, and governable — a property to be priced, bounded, and engineered against, like corrosion, like side effects, like every cost that came with every capability civilization has ever adopted. The chapters that follow build the governance, component by component, and the reader who finishes them will never again meet a fluent falsehood without knowing exactly what it is, where it came from, and what to do about it.

- [ ] The organization's terminology distinguishes hallucination from error, uncertainty, and deception. - [ ] The five-family taxonomy (fabrication, attribution, overgeneralization, instruction drift, inheritance) is applied in risk registers. - [ ] Each family is mapped to its appropriate machinery (grounding, provenance, calibration, control, curation). - [ ] Neither the shrug nor the panic posture appears in training or communications.

The word has a history worth reading, because the history shows a phenomenon outrunning its vocabulary. The clinical term — hallucination — belonged to perception: seeing what is not there, in the senses' own theater. The field borrowed it for its systems because the public experience rhymed: a machine reporting on a world it was not perceiving. The research community quickly reached for the cognitive alternative — confabulation — whose clinical meaning (fluent gap-filling without intent to deceive) fit the mechanism precisely, and the two words have coexisted since, doing different jobs: the borrowed word for the public's experience, the clinical word for the engineers' mechanism, and a steady trickle of papers proposing quieter terms — fabrication, ungrounded generation, deviation — none of which has displaced the pair, precisely because the pair already splits the phenomenon along the book's own line: what it feels like from outside, what it is from inside.

The vocabulary's history teaches a second lesson the reader should keep: definitional drift is the phenomenon's early symptom. In public speech, "hallucination" has begun to mean any machine error, any bad output, any answer the speaker disliked — a drift that matters because it destroys the

diagnosis the taxonomy requires: the program that files every failure as hallucination buys one tool (or one apology) for five families of failure, and the reader of Chapter 1's taxonomy knows what that costs. The discipline's response is the same as every technical field whose word escaped into the streets: define precisely inside the program, tolerate the blur outside, and never let the blurred public meaning write the internal risk register.

The diagnostic for this section is one sentence: ask the organization what the difference is between a hallucination and an error. A structured answer (the tether, the mechanism, the family) is a program that has kept its vocabulary; a shrug is an organization about to misdiagnose its failures in public, which is where the word learned its blurred meaning in the first place.

The phenomenon's danger is not only in the machine; it is in the meeting, and the chapter's second section reads the meeting's psychology, because the reader's mind is the falsehood's actual target. The cognitive traditions established the relevant effects long before the machines: processing fluency — the ease with which a statement is read — is registered by readers as a signal of truth; the polished sentence feels truer than the clumsy one, independent of content. The machine is the most fluent writer the population has ever met: grammatical, structured, confident, tireless — every surface cue that human readers evolved to trust, produced at zero cost, about content with any truth value.

The illusion has a second layer: the coherence effect — statements woven into structure (lists, headings, citations) are judged more credible than the same claims scattered — and the machine's outputs are nothing but weave. And a third: the consistency bias — a claim encountered in confident prose reads as familiar, and familiarity is read as verification: the reader who has seen the fabricated citation once will, seeing it again in the machine's next answer, experience the quiet click of "that checks out," from nothing but repetition.

The reader should see the composition: a generator that produces the perfect surface for any content, meeting an audience whose truth-judgments are substantially surface-read. This is not an accusation of the audience; it is the operating condition. Human skepticism was calibrated on adversaries — writers with incentives — and its heuristics (fluency as competence, confidence as authority, structure as care) were cheap and mostly right, because producing the surface used to cost effort and intent. The machine removed the cost and the intent while keeping the surface, which breaks the heuristics' economics: the cues are now free, so the cues no longer price anything.

The program's consequence, and the whole of Chapter 6, is that the defenses

must move outward — into checking the content, by machinery and by trained habit — because the surface has been captured, permanently, by the blur. The reader who understands this section has stopped asking why people believe the machine, and started asking what would make belief expensive again, which is the right question, and one this book spends nine more chapters answering.

The chapter closes its definition work at the boundary, because the phenomenon's edge cases are where programs argue and disciplines mature. The first edge: fiction. A story is untethered output by design, and the same generator produces it; hallucination is not the word for the novel any more than for the author's. The boundary is the frame — the output presented as about the world is the phenomenon's home; the output presented as invented is not — and the boundary's failure mode is frame collapse: the fiction that leaks into factual consumption (the machine's invented example quoted as a case study; the hypothetical attributed as an event), which is the attribution family wearing fiction's clothes.

The second edge: prediction. The forecast is untethered by nature — the future has no corpus — and the disciplined forecast states its grounds and its uncertainty; the confabulated forecast states its fiction with a forecast's confidence. The boundary, again, is the claim's relationship to checkable grounds, not the claim's truth: forecasts can be wrong without being hallucinated, and hallucinated forecasts can be right by luck. The third edge: the plausible-but-unverifiable — the output about a private event, an oral tradition, an unindexed reality. The reader cannot check it; the machine cannot ground it; and the honest program marks the category rather than resolving it: untethered-and-uncertifiable, an epistemic no-man's-land where the book's disciplines can only require the frame to say so.

The reader should note what the three edges share: in each, the phenomenon's defining property — the tether, not the truth — is the discriminator. The program's vocabulary work is done when the edge cases stop being arguments and become classifications: a proposed output can be asked which family it is, what frame it claims, and what grounds it has, and the answers route it — fiction to the frame discipline, prediction to the calibration discipline, the unverifiable to the disclosure discipline, and the unframed ungrounded-about-the-world to the machinery of the rest of this book. The organization whose reviews can run that routing in one question has a working taxonomy; the organization still arguing about words has not yet started engineering, and its risk register will show it — every failure filed under one name, every fix bought once and expected everywhere, which is the theater's accounting, applied to epistemology.

The chapter's first practical section interrupts a habit the reader has probably already formed, because the interruption is the taxonomy's most useful application: the question "what is the hallucination rate?" — the industry's headline number, the launch deck's centerpiece — is the wrong first question, and the reader should know why in structural terms. The question presumes the phenomenon is one thing (Chapter 1's first correction: five families, five machineries); it presumes a uniform population (Chapter 2's geography: the rate is a function of the corpus's density, and one number averages the famous with the obscure into a fiction that matches neither); and it presumes the rate is the decision (Chapter 3's costs: the rate in the ambient domain and the rate on the record are the same number with prices that differ by orders of magnitude — the rate that matters is the rate in the domain the deployment touches, weighted by what a falsehood there costs).

The correct first question, the one the program should ask before any rate is measured: which families, which regions, which domains, and what does each failure cost there? — the taxonomy's four-part walkthrough, which produces a risk register instead of a number, and a risk register is what everything downstream (the tiers, the machinery, the measurement) is actually built from.

The section's warning is about the number's gravity: the single rate is not merely uninformative; it is governance-corrupting, because it gives the organization a metric to manage that optimizes for the appearance of the wrong thing — the score improving while the failures migrate to the load-bearing domains, the suite's shape learned while the frontier's behavior is unmeasured, the theater of Chapter 8 arriving before the measurement has even started. The mature program treats the headline rate the way it treats every summary: as a derived, domain-weighted roll-up of the register, produced after the real questions are answered, never instead of them.

The diagnostic: when the organization reports its hallucination number, ask for the register it was rolled up from. The register exists in a real program, and the number is its honest shadow; the number alone is the blur's last performance

— a confident, fluent, specific figure, untethered from any machinery that produced it, which the reader will notice is exactly the phenomenon this book is about, arriving as its own statistic.

The taxonomy chapter's second section widens the vocabulary lens, because the phenomenon is global and its governance will be cross-cultural, and the reader who builds or deploys across languages meets a fact the single-language discussion misses: the confabulation's geography is different in every language, and so is its cost. The mechanism chapter's geography applies per language: the corpus's density in a language determines the

machine's fluency-truth correlation in that language — the heavily-sampled languages inherit the machine's best regions, and the sparsely-sampled ones inherit the blur at its worst, which means the global deployment's hallucination risk is not one risk but a map of them: the same system that grounds well in its oversampled languages confabulating fluently in its undersampled ones, on populations who receive the output with no surface-level way to tell the difference — the defense asymmetry of Chapter 3, cut along the world's language lines, with the least-defended populations again paying the most.

The vocabulary layer matters too, and the section names its stakes: the epistemic norms of checking differ across traditions (the citation cultures, the oral-record traditions, the trust structures of institutions), and the machine's fluent confidence lands differently in each — in a culture whose norms defer to written authority, the confabulated citation inherits a deference the machine did not earn; in one whose norms prize the elder's verification, the machine's confident prose may be discounted appropriately by habit the program did not build and cannot claim. The program's discipline is the book's standing one, applied across languages: the coverage map per language (the corpus chapter's map, drawn per language, with the sparse-language regions marked as elevated risk rather than silently served); the measurement sliced per language (the suite of Chapter 8 built in each deployment language, not translated from the dominant one — because the abstention grammar of the suite is itself a language-specific behavior); and the disclosure honest about the differences (the four-clause sentence per language region, because a single global rate over a multilingual deployment is the wrong first question of the previous section, multiplied across every line the deployment speaks).

The diagnostic: ask any multilingual deployment for its rate in its least-sampled language. The answer's existence — measured or honestly unknown — is the finding; and the reader can guess which answer the launch deck contains, in every language it ships.

The chapter closes by making the taxonomy operational — the walkthrough a design review actually runs, in the form a program can adopt. The setup: a proposed deployment, a room, and the five families as the agenda. Fabrication: the review asks what specific entities the system will state — citations, identifiers, numbers, names — and for each, what tethers it: the corpus that contains them, the lookup that verifies them, or nothing; the "nothing" answers become the register's fabrication rows, and the reader who wants the review's signature question has it: where does that number come from, at runtime?

Attribution: the review asks who and when the system binds facts to — the

binding machinery (retrieved with the fact, or generated beside it), and the attribution rows record the bindings that are composed rather than fetched. Overgeneralization: the review asks where the system states rules — the populations and contexts the rules will reach, and the exceptions the training averaged away; the rows record the domains where the average is the harm. Drift: the review asks what the system's instructions constrain and what enforces the constraints' persistence across a long session, a changed context, an adversarial user — the drift rows record the constraints enforced by hope.

Inheritance: the review asks what the corpus contained — the known falsehoods, the stale facts, the contested claims — and who curates them out or marks them; the rows record the corpus's known lies.

The walkthrough's yield is the register, and the register's shape is the review's real lesson: the five families do not produce five problems; they produce a map of the deployment's specific confabulation surface, with each region owned by a named machinery (Chapter 1's mapping) or honestly marked as carried residual — and the program's maturity is readable from the register's bottom rows: the residuals, named, priced, signed. The walkthrough takes an hour, costs a room, and outperforms every rate-measurement the organization could run first, because it prices the geography before the sampling and names the machinery before the incident — which is the order the book has argued for from its first page, and the order this review makes routine.

The diagnostic: ask any deployment for its five-family register. The register that exists was built by a program; the register that would need a meeting to construct is the program the organization plans to have after the incident, which is the chapter's last echo of the entire series: the machinery that exists before it is needed is the only machinery that is ever there when it is.

Hallucination is output wearing knowledge's clothes without knowledge's tether: fluent, confident, specific, and about a world the system is not actually consulting. Beneath the single word sit five distinct failures — fabricated specifics, wrong attributions, overreached rules, drifted instructions, inherited falsehoods — each demanding a different engineering. The phenomenon is neither a quirk to be enjoyed nor a reason to abandon the machinery; it is the shadow of the capability, priced and governed like every shadow that ever came with power. The next chapter opens the machine's case to show why the shadow falls at all — and why the obvious fix is not a fix.

End of free sample. The full book is available at putrabooks.github.io

ABOUT THE AUTHOR

Aiman JoeHanz writes the Building of Tomorrow series, a volume-per-technology account of the systems defining this decade -- from the algorithms and their ethics to the cities, medicine, and stations they will build. Each book is a structural treatment: the parts, the failure modes, the economics, and the program. In this volume the author takes up the machine era's signature flaw -- the fluent, confident falsehood -- and treats it the way the series treats every problem: as engineering, with a mechanism, a cost, a defense in depth, and a calendar to run it on.