SAMPLE

Building the AI Ethics & Safety of Tomorrow

Building the Trust Discipline 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-ethics-safety-of-tomorrow-ebook-pdf-1465661239

Ko-fi

ElevenReader

Gumroad

Payhip

Contents

Introduction

Chapter 1: The Discipline's Anatomy

Chapter 2: The Harm Taxonomy

Chapter 3: The Specification Problem

Chapter 4: The Evaluation Problem

Chapter 5: The Adversarial Discipline

Chapter 6: Transparency and Explanation

Chapter 7: Privacy and Data Dignity

Chapter 8: Governance

Chapter 9: Deployment and the Safety Case in Production

Chapter 10: The Operating Discipline

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

Introduction

Every technology that acts on people acquires, sooner or later, a safety culture. Boilers begat inspection codes. Automobiles begat crash standards. Medicines begat trials. Aviation begat checklists so effective that a safety failure now makes international news precisely because it is rare. The machines that make decisions — that recommend, rank, screen, price, approve, and increasingly act — are acquiring theirs now, and this book is about the discipline taking shape: the ethics and safety of artificial intelligence, treated the way this series treats every technology, as an engineering practice that can be described structurally, built deliberately, and operated on a calendar.

The subject attracts two kinds of writing, both inadequate for the practitioner. The first is moral philosophy at altitude: principlism — lists of high values like fairness, transparency, and accountability — that everyone endorses and no one can deploy, because a value statement does not tell an engineer what to test on Tuesday. The second is the technical narrow-cut: benchmark chasing, red-team stunts, compliance checklists sold as governance. This book takes the middle path the series has taken for every other technology: the structural treatment. What are the parts of the discipline? What does each part actually do? How do the parts fail, and what does a failure cost? And what would a working ethics-and-safety program — the kind an organization can run, audit, and improve — look like as an engineered system?

The argument of the book can be stated in three claims. First, ethics and safety are not a mood or a document; they are controls — specific mechanisms that reduce specific risks to specific people, and a principle that cannot be turned into a mechanism is not yet operational. Second, the discipline's center of gravity is the lifecycle: a machine system acquires its ethical properties the same way it acquires its other properties, through design choices, test regimes, deployment boundaries, and monitoring — made by identifiable people on identifiable dates, or not made at all. Third, the arrival of more capable and more agentic machine systems does not change the discipline's structure; it raises its stakes, in exactly the ways this book will price chapter by chapter.

Ten chapters follow. The first three build the foundation: what ethics and safety distinguish from each other, how machine systems actually harm people (a taxonomy, because "harm" is too vague to engineer against), and the specification problem — the oldest and hardest question in the field, which is whether the instructions a machine follows can ever mean what the instructing organization intended. Chapters four and five cover the discipline's

measurement apparatus: evaluation, the problem of testing systems whose behavior cannot be fully enumerated, and red-teaming, the adversarial half of the same practice. Chapters six and seven treat the two disciplines most organizations meet first: explanation and interpretability, and privacy.

Chapters eight and nine move from mechanisms to programs: governance, the boards and policies and audits that turn controls into an institution, and deployment, the place where all of it either operates or was theater. The tenth chapter assembles the whole into an operating discipline — the program an organization can actually run — and closes with the questions this series always leaves with its readers, sharpened for the technology that makes decisions.

A note on method, consistent with the series: this book is evergreen. It names no products, cites no benchmark scores, and quotes no market statistics, because all three expire; the structures — the risk taxonomies, the control patterns, the program anatomy — are the layer that lasts. Vendors are named by category. Claims are hedged where the field's knowledge is genuinely unsettled. And every chapter ends the same way: with a closing argument, a checklist, and a diagnostic — because a discipline that cannot be inspected is a slogan, and this book is not a slogan.

Chapter 1: The Discipline's Anatomy

The field's two words are often used interchangeably, and the first act of a working discipline is to take them apart. Safety, in the engineering tradition, is about preventing states of the world that are bad in ways almost everyone agrees on: injury, death, financial ruin, the exposure of private life, the loss of rights. Its questions are operational — bad according to whom, how bad, how likely, and what reduces the likelihood — and its answers are mechanisms. Ethics is the larger container: the discipline of asking what ought to be the case, which includes the questions safety answers but also the genuinely contested ones — fairness in allocation, the boundaries of permissible influence, the respect owed to persons as decision-makers rather than decision-consumers.

The working relationship between the two is asymmetric and productive: ethics sets the values and identifies the conflicts; safety turns the agreed parts into controls; and the un-agreed parts are governed by a different machinery — deliberation, policy, disclosure — rather than being smuggled into a test suite.

A machine system can be ethical without being safe, and safe without being ethical, and both combinations teach the reader the distinction. A recommendation system can be perfectly safe — it injures no one, leaks nothing, respects every privacy law — while being ethically corrosive, because its optimizing objective captures attention in ways its own designers find regrettable when named plainly. And a hiring screen can be built with the most careful ethical deliberation about fairness, yet remain unsafe in the engineering sense if it fails unpredictably at the volume of a large employer. The discipline's first structural insight is that neither word can be delegated to the other: the safety function cannot decide, alone, what fairness requires; and the ethics function cannot ship anything, at all, without the safety function's mechanisms.

Every working ethics-and-safety practice, examined structurally, contains five components, and an organization can be diagnosed by which components it possesses.

The first is risk identification — the standing inventory of the ways the machine can make things worse, for whom, and through what mechanism. The second is specification — the translation of intended behavior into instructions precise enough to build against: objectives, constraints, and the definition of good that the system will actually optimize. The third is assurance — the apparatus that produces evidence the system behaves as specified: tests,

evaluations, audits, adversarial probes. The fourth is control — the runtime machinery that bounds what the system can do: permissions, thresholds, human checkpoints, the ability to slow or stop. The fifth is accountability — the assignment of every one of the former to a person with authority and budget, because an unowned control decays into a document.

The components are coupled, and the couplings are where the real work lives. Risk identification without specification produces anxiety without direction. Specification without assurance produces intentions with no evidence. Assurance without control produces beautifully documented systems that misbehave in production anyway. And all four without accountability produce the field's most common artifact: the ethics program that exists to be photographed — the published principles, the annual review, the absence of anyone whose job performance depends on a control working.

Structurally, an organization's ethics-and-safety function sits in an uncomfortable position, and the position itself is informative. The function must be close enough to the builders to see the actual system — its data, its objective, its deployment surface — or it will govern a fiction. It must be far enough from the delivery pressure to say no, or it will govern nothing. The mature placement borrows from the professions the discipline admires: the safety function reports through a line that does not depend on the delivery's success, the way audit reports through the audit committee and aviation safety reports outside the flight schedule; it keeps its own evidence trail; and it exchanges a veto power for something more durable — the standing right to publish its findings inside the organization, which is the power that survives the first budget crunch.

The immature placements are familiar to anyone who has worked in a technology organization. Ethics as marketing — owned by communications, producing principles and little else. Ethics as compliance — owned by legal, producing the minimum the local regulation demands, measured in documents rather than controls. Ethics as heroics — no function at all, but a reputational engineer who appears after incidents to absorb blame. Each placement is diagnostic: where the discipline sits in the org chart is what the organization actually believes it is for.

The series' reader is equipped, throughout, with diagnostic questions — the one-minute interrogation that reveals an organization's actual maturity. This chapter contributes the first: ask an organization to name the difference between its ethics principles and its safety controls. A mature answer names artifacts: these principles, these controls, this mapping between them, this owner for each. An immature answer recites values — or, worse, recites the

difference between the words as if the question were etymological. The distinction this chapter has drawn is not academic; it is the difference between a discipline that produces mechanisms and one that produces literature, and the chapters that follow are concerned entirely with mechanisms.

- [ ] The organization's ethics values are each mapped to at least one named control or governed process. - [ ] A document distinguishes safety risks (agreed harms) from ethical questions (contested values) and assigns each a different machinery. - [ ] The five components — risk identification, specification, assurance, control, accountability — each have a named owner. - [ ] The ethics-and-safety function's reporting line does not depend on delivery success. - [ ] The function has the standing right to publish its findings internally.

The discipline is young, but its short history has a shape worth reading structurally, because the shape recurs in the maturation of every safety field. The first act was the reckoning: a small number of widely-discussed deployments — in criminal justice, in hiring, in credit — forced the recognition that optimization systems reproduce the patterns in their data, and the first wave of work was diagnostic, cataloging the failure patterns this book calls allocation and quality harms. The second act was the safety turn: as systems grew more capable and more general, the field's attention widened from discrimination to the broader questions of what these systems do under pressure, at scale, and in the hands of adversaries — the evaluation, red-teaming, and specification machinery this book's middle chapters describe. The third act, still in progress, is institutionalization: the movement of the discipline from research papers into review boards, regulation, procurement requirements, and the audit traditions — the machinery of Chapters 8 and 9.

The reader should notice what the history teaches about the field's present: the acts overlap rather than replace each other, and an organization today can be in any of them — still arguing about whether its systems should be reviewed at all (act one), building its evaluation practice (act two), or running governance (act three). The diagnostic value is real: ask which act an organization's ethics program is living in, and its next two years are largely predictable, because the acts' transitions are the field's reliable pattern — each one forced by a class of incident, and each one cheaper to enter voluntarily than by the same door.

The discipline is spoken by three communities whose languages differ enough that their mutual misunderstandings are themselves a hazard worth mapping. The research community speaks in formalisms: metrics, bounds, mechanisms — its strength is precision, its temptation is the toy problem, the formalization

that captures the clean part of a messy world. The practitioner community speaks in pipelines and incidents: its strength is contact with reality, its temptation is the checklist that substitutes for judgment — the maturity to know that a control ran, without asking whether the control works. The policy community speaks in duties and powers: its strength is scale and enforceability, its temptation is the mandate that specifies the document rather than the outcome, regulating the artifact instead of the harm.

The mature program is trilingual, and the translation work is a named duty, not a hope: research findings become evaluation instruments; engineering practice becomes standards; policy becomes checklists — and, in the upward direction this book insists on, incidents become findings, findings become policy. The reader who serves on a review board will spend a surprising fraction of every meeting translating: the metric the engineer cites, the duty the lawyer cites, and the harm the community member describes are often the same fact in three dialects, and the board that cannot translate will decide on the dialect it knows rather than the fact all three describe. The field's best programs make the translation explicit — a standing glossary, a shared register, a culture that rewards the person who asks the question that converts one language into another before the decision, rather than after.

Before the book builds its maturity model, it should demolish the field's three most common false measures. Maturity is not principles: the organization with the longest values list is frequently the one with the fewest controls, because the list is the program — writing values is cheap enough to substitute for building machinery, and frequently does. Maturity is not tools: a portfolio of evaluation dashboards and monitoring software measures procurement, not practice; the question is never what instruments exist but what decisions they changed. And maturity is not headcount: a large function without authority is a cost center that will be right-sized at the first budget crunch, and the safety tradition's experience is uniform on this point — the function's size matters far less than its reporting line, its evidence trail, and its right to publish findings.

What maturity is, this book has already begun building: artifacts that can be inspected, decisions that can be traced to owners, controls that are tested rather than presumed, and the institutional memory that turns incidents into policy. The ladder the discipline can honestly offer runs, roughly: absent (nothing), reactive (responses after incidents), documented (principles and policies without the machinery), operational (the machinery exists and runs on a calendar), and learning (the machinery feeds itself — every incident and finding visibly changes the register, the tests, and the policy). The reader should carry the ladder as this section's deliverable, and the diagnostic it generates: ask any program for the change its last incident produced, and the

answer places the program on the ladder more accurately than any self-assessment.

The book's chapters describe the program; this section names the builder's personal share of it, because the discipline is ultimately distributed across the people who make design choices, and four responsibilities fall to them specifically. The first is disclosure: the builder is the only person positioned to know what the system actually does — its data, its objective, its failure patterns — and the discipline begins with their honest account of it, in the specification and the documentation, rather than the version the launch wants. The second is refusal: the designer's standing, and protected, ability to decline the instruction whose specification they cannot defend — the professional tradition's oldest right, imported from engineering's codes, and the organizational duty to make the refusal survivable, because a refusal that ends careers ends nothing else.

The third is escalation: the obligation to carry concerns past the team, through the machinery of Chapter 8, when the team will not act. And the fourth is craft: the habit of design that builds the ethical questions in from the start — the boundary drawn before the integration, the disaggregation planned before the dashboard — because questions asked at specification cost a meeting and questions asked at production cost a program.

The builder's responsibilities are not the program's substitute; they are its material. The mature organization treats them as a professional standard, taught in the formation, protected in the escalation machinery, and rewarded in the review — and the reader who builds machine systems should carry the four as their personal register, the way the older safety professions carry theirs: quietly, permanently, and as the identity of the work rather than the compliance around it.

One word hovers over the discipline that this book has deliberately kept out of its component list, because it is not a component: trust. Trust is the discipline's output — the property an organization earns by running the machinery visibly and losing it by running the theater — and its structural properties deserve a paragraph. Trust is earned in the small and lost in the large: the appeal answered well, the finding published honestly, the launch delayed credibly — these accumulate; and a single incident handled defensively erases years of accumulation, because the public's model of an institution is updated most sharply by its worst moment. Trust is also asymmetric across audiences: the regulator's trust is built on evidence, the user's on experience, the expert community's on candor about limitations — and the theater fails each audience differently, being detectable by each.

The reader should resist the field's most common trust move: trust as marketing, the safety program run primarily for its publicity value. The pattern is self-defeating in a way the marketing instinct never prices: trust built by claims is spent by incidents, and the claims raise the incident's cost — the organization that advertised its ethics most confidently owns the largest credibility gap when the harm surfaces. The mature position inverts the sequencing: build the machinery, let the trust follow, and treat the publicity as a lagging indicator rather than a goal. The diagnostic for the section: compare an organization's ethics communications to its published findings. Communications exceeding findings is the theater's signature — and the reader who has followed this book can now read the ratio at a glance.

An honest harm taxonomy must include the harms of refusal, because the discipline is regularly abused in exactly one direction: systems that would reduce harm — earlier detection, broader access, cheaper expertise — delayed, diluted, or prevented by programs whose real function is institutional self-protection. The safety function, like every control function before it, can become a veto without a docket: no decision record, no proportionality, no cost attached to saying no — and the accumulated cost of the noes is invisible precisely because prevention produces no incidents. The reader should hold the symmetry plainly: a wrongful deployment harms the population it touches; a wrongful refusal harms the population it never reaches, and the second population does not appear in the incident reports, the complaint queue, or the news.

The mature program prices both directions. Its review asks not only what the harm of launching is, but what the harm of not launching is — the differential that made the older safety fields insist on proportionality rather than prohibition — and its decisions are recorded with both prices, so the no is accountable the same way the yes is. The immature program inverts this: the no is free, the yes is expensive, and the institution drifts toward the posture that costs its reviewers nothing and the public everything. The diagnostic for the section asks any program for its last refusal and its price: what was forgone, who bore the forgoing, and where is the record? An answer exists in a working program, because both errors are errors — and the discipline that cannot see one of them is only half a discipline, whichever half it prides itself on.

The anatomy chapter closes with the question every control function eventually faces from its funders: how do you know the program works? The safety tradition's honest answer is the leading-indicator problem, and it deserves the structural treatment. The program's success metric — harms that did not happen — is unobservable, which leaves the observable proxies:

findings found (the register's count, which rises with honesty rather than with danger), controls tested (the audit's pass rate on instruments, not documents), appeals resolved and reversals made (the accountability machinery's throughput), escalations raised (the reporting culture's pulse), and the time-to-correction of the incident record (the learning loop's speed). Each proxy is gameable — a register padded with trivia, an audit graded on its own rubric — and the program's measurement discipline is the same one it applies everywhere else: multiple instruments, disaggregation by stakes, and the qualitative channel read as a first-class sensor.

The reader should note the measurement problem's deeper form: the program that succeeds produces silence, and silence is indistinguishable, from inside, from the program that does nothing — the pattern the security tradition knows as the paradox of the uneventful year. The mature response is the demonstration rather than the claim: exercises (the incident rehearsed, the appeal audited, the red team's finding surfaced and closed in public view), published to the internal record so the silence is documented as the program's product. The immature program faces the same silence and produces, instead, the activity report — the trainings delivered, the meetings held, the policies published — the theater's measurement, counting the performance rather than the protection. The diagnostic asks for the difference: what changed because the program existed? One example, in a working program, is a control that was built; a page of activities, in the theater, is the performance — and the reader, by now, can tell the two apart in a single reading.

The anatomy is completed by its boundaries — the neighboring disciplines the ethics-and-safety function works beside, trades with, and must not be confused with. Legal and compliance: the lawyers own the obligation's letter and the regulatory relationship; the safety function owns the machinery beneath it — and the two fail together when one substitutes for the other, the compliance posture governing where the risk needed engineering. Security: the closest sibling, sharing the adversary's worldview and the incident machinery; the division is one of focus — security protects the system and its data from misuse, ethics-safety protects people from the system's own intended operation — and the deployments need both, on the same register.

Product: the discipline's permanent counterparty and essential partner — the function that knows what users actually do, and the owner of the surfaces where the harms of Chapter 2 arrive. And the domain functions — the clinicians, the credit officers, the moderators, the educators — whose professional judgment the machine encodes and whose stakes the program must represent in every review.

The neighbor map matters for two structural reasons. First, the handoffs: every boundary in the map is a seam where a risk can fall between owners — the security finding that is also a safety harm, the product change that quietly widens a boundary, the legal opinion that becomes, by default, the ethics decision. The mature program names the seams (the shared register, the joint reviews, the escalation path that routes by risk rather than by department). Second, the vocabulary: each neighbor speaks the discipline's facts in its own dialect, and the reader who learned the three communities' languages in the earlier section can generalize — the program's practitioners translate, or the seams stay quiet, and the quiet seam is where the incident's causal chain begins.

The diagnostic asks for the seams: where do the safety function and its neighbors meet, and what is written at each meeting? Named joints are a working anatomy; the empty space between departments is the shadow program, again — the risk, ungoverned, precisely where everyone assumed someone else was holding it.

Ethics and safety, for machine intelligence, are not a sentiment; they are an engineering discipline with five load-bearing parts — what can go wrong, what "right" means, how we know, how we bound it, and who answers for it. The parts are separable, diagnosable, and buildable, and an organization can be read like a blueprint by which parts it has actually built. The chapters that follow take the parts one at a time, because the discipline's hardest problems — specification, evaluation, the economics of assurance — deserve the structural treatment the series gives every technology it examines: anatomy first, economics second, failure modes always, and the operating discipline last.

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

ABOUT THE AUTHOR

Aiman JoeHanz is the author of the Building of Tomorrow series, a technology-by-technology account of the systems arriving in this decade -- from blockchain and machine intelligence to energy, space, and the ethical machinery that governs them. Each volume is written the same way: no vendor names, no benchmark scores, no forecasts pretending to be facts, only the structure that lasts -- the parts, the failure modes, the design choices, and the program a reader could actually run. In this volume the author turns that method on the discipline itself, treating ethics and safety not as a sentiment but as the load-bearing layer beneath every other machine the series builds.