F

FAIR LABS

FAIR ARTIFICIAL INTELLIGENCE RESEARCH LABS

GUIDELINES ◆ AI POLICY ◆ 2025

Ethical AI Development Guidelines

A practical program for building artificial intelligence responsibly—translating ethical principles into disciplined process, ethical review, and genuine stakeholder engagement.

PROGRAM	DOCUMENT TYPE	AUDIENCE	PUBLISHED
AI Policy	Guidelines	Research & Industry Teams	2025

FAIR Labs · A 501(c)(3) nonprofit research institute · Washington, DC

CONTENTS

Table of Contents

EXECUTIVE SUMMARY

From principles to practice

IN ONE PARAGRAPH

Ethical AI is not a statement of values pinned to a wall; it is a set of practices woven into how a team frames problems, gathers data, engages the people it affects, reviews consequential decisions, and anticipates misuse. These guidelines convert five widely shared principles into operational obligations across the development lifecycle, propose a proportionate ethical review process adapted from human-subjects research, and confront directly the failure mode that discredits the whole enterprise—ethics performed for appearance rather than practiced for effect. The instruments here are meant to be lifted into a team's actual workflow.

5
core principles that anchor every downstream obligation

6
lifecycle stages, each with ethical duties and a decision artifact

3
review tiers that scale scrutiny to the stakes of the work

What we recommend

1. **Make ethics a process, not a posture.**
Give every principle an owner, a trigger, and an artifact. An ethical commitment that leaves no inspectable trace is indistinguishable from no commitment at all.

2. **Engage the affected before you build.**
Treat the people a system will touch as design partners, not as focus groups consulted after the architecture is fixed. Participation early is cheaper and more honest than remediation late.

3. **Establish proportionate ethical review.**
Stand up an IRB-like review with clear triggers, a competent and independent board, and written determinations. Scale the depth of review to the consequences of the work.

4

Anticipate misuse, not just error.

Ask not only whether a system works as intended but how it will be turned to purposes you did not intend. Dual-use foresight belongs in design, release, and monitoring—not in hindsight.

The remainder of this report develops each recommendation into operational detail, with mappings, triggers, and review procedures a team can adopt directly.

SECTION I

Why Ethics Needs Process, Not Just Intentions

Almost no one sets out to build a harmful system. The engineers who trained a recruiting model that learned to penalize women's resumes wanted a fair tool; the teams behind facial analysis systems that failed on darker-skinned faces were not indifferent to accuracy; the designers of recommendation engines that amplified outrage were pursuing engagement, not radicalization. Harm in AI rarely originates in malice. It originates in the ordinary momentum of building—decisions made quickly, under deadline, by people who never asked a question that would have surfaced the problem while it was still cheap to fix. This is the central premise of these guidelines: ethics fails not for want of good intentions but for want of a process that makes intentions consequential.

Intentions have three properties that make them unreliable as the sole safeguard. They are *private*, so an organization cannot inspect them, audit them, or hold anyone to them. They are *fragile*, evaporating under the pressure of a shipping date or a revenue target. And they do not *scale*: a founding team's care does not automatically transmit to the fiftieth engineer or the third acquiring company. Process addresses all three. It externalizes commitments into artifacts that can be examined, it embeds them in gates that a deadline cannot simply bypass, and it transmits them to newcomers who were not present when the values were formed.

The limits of the ethics statement

Most organizations that build AI have published a set of ethical principles. Very few can show what those principles changed. The gap between the principles a company professes and the decisions it actually makes is the space in which "ethics washing" grows—the performance of ethical concern in place of its practice. A principle that never blocks a launch, never reshapes a dataset, and never costs anything is not a principle; it is a decoration. The test of an ethics program is not the elegance of its statement but the record of times it changed an outcome.

A WORKING DEFINITION

Throughout this report, **ethical AI development** means the disciplined practice of anticipating, surfacing, and addressing the human consequences of a system throughout its life—not a property a finished model possesses, but a quality of the process that produced and maintains it. The unit we govern is the *practice*, not the artifact.

What process buys that principles alone cannot

A well-designed ethical process does four things a statement cannot. It **surfaces questions at the moment they are cheap to answer**—during framing and data collection, rather than after deployment

when remediation means recalls and apologies. It **distributes the cognitive load** of foresight across a team and a review body, so that no single person must anticipate every failure alone. It **creates a record** that lets an organization learn from its own decisions and demonstrate diligence to those it affects. And it **assigns ownership**, converting a diffuse sense that "someone should think about this" into a specific person who is answerable.

None of this requires an organization to become slower or more timid. The purpose of process is not to obstruct; it is to make the many low-stakes decisions frictionless precisely because the few high-stakes ones receive real attention. A team that reviews everything reviews nothing seriously. Proportionality—doing more where the stakes are higher and less where they are lower—is what makes an ethics program affordable enough to be genuine.

> *Good intentions are not a control. They cannot be audited, they do not survive a deadline, and they do not transfer to the next engineer.*
>
> — The governing premise of these guidelines

Who these guidelines are for

We write for practitioners: the researcher deciding whether to release a model, the engineer choosing a training corpus, the product manager scoping a feature, the executive setting the incentives that will decide whether any of the foregoing has time to think. Ethics is often addressed to none of these people—pitched instead to philosophers or to regulators. That is a mistake. The decisions that determine whether a system harms people are made by builders, in the course of building. These guidelines meet them there, and everything that follows is written to be used rather than merely admired.

SECTION II

Core Principles for Ethical AI

Principles are not the destination; they are the compass. Five of them, drawn from bioethics and refined for AI, are broad enough to endure and concrete enough to translate into the obligations that occupy the rest of this report.

The AI ethics literature has produced a proliferation of principle sets—dozens of frameworks from companies, governments, and standards bodies. Encouragingly, they converge. Beneath the varying vocabulary lies a stable core, and that core maps cleanly onto the four classical principles of biomedical ethics—beneficence, non-maleficence, autonomy, and justice—supplemented by a fifth that the opacity of machine learning makes newly urgent: explicability. We adopt these five not because they are novel but because they are tested, mutually reinforcing, and each independently actionable.

The five principles

Each principle answers a distinct ethical question. Together they form a nearly complete inventory of the ways an AI system can go right or wrong for the people it touches.

THE FIVE CORE PRINCIPLES AND THE OBLIGATION EACH IMPOSES

Principle	The question it forces	The obligation it imposes on builders
Beneficence	Does this system do genuine good for real people, and for whom?	Articulate the benefit, name the beneficiary, and verify the good is delivered—not merely intended.
Non-maleficence	What harm could this cause, to whom, and how likely is it?	Anticipate and mitigate foreseeable harm; when in doubt, favor caution over speed.
Autonomy	Does this respect people's capacity to decide for themselves?	Preserve meaningful human choice, informed consent, and the right to contest.
Justice	Are the benefits and burdens distributed fairly across groups?	Test for disparate impact; ensure the vulnerable are not made to bear the costs.
Explicability	Can we understand and account for what the system does and why?	Build for interpretability and transparency; be able to explain and to answer for outcomes.

Beneficence and non-maleficence: the two-sided ledger

Beneficence asks a question that developers skip with surprising frequency: what good, precisely, does this system do, and for whom? Much AI is built because it is possible, or because a competitor built something similar, without a clear account of the benefit or the beneficiary. Naming the good

concretely is disciplining—it reveals when the true beneficiary is the builder rather than the user, and it establishes a standard against which the delivered benefit can later be checked. Non-maleficence is beneficence's necessary complement. The injunction to "first, do no harm" carries particular force in AI because harms are frequently distributed, delayed, and invisible to their authors—a subtle bias felt only by a minority, a manipulation that works below awareness, a dependency that erodes a skill over years. The obligation is to look for these harms deliberately, because they do not announce themselves.

Autonomy: preserving human agency

Autonomy is the principle most easily eroded by systems built to be helpful. A recommendation engine that narrows what a person sees, a default that nudges a choice, an automated decision that offers no route to a human—each can diminish agency while appearing to serve it. Respecting autonomy means preserving the conditions under which people can decide for themselves: genuine choice rather than manufactured consent, transparency about when a machine is acting, and a meaningful ability to opt out or to contest. The principle is not hostile to automation; it is hostile to automation that quietly removes the human from decisions that are properly theirs.

DESIGN RULE

When two principles conflict—as beneficence and autonomy often do, when a system could help people by deciding for them—the resolution is not to pick a winner in the abstract but to surface the tension explicitly, document the trade-off, and let an accountable human make and record the call. Ethics is frequently the management of genuine tension, not the application of a formula.

Justice: fairness in distribution

Justice concerns how a system's benefits and burdens fall across people. An AI system can be accurate on average and unjust in distribution—performing well for the majority while failing the groups least equipped to complain. Because models learn from historical data, they inherit historical injustice unless deliberately corrected; a system trained on past decisions will faithfully reproduce the biases embedded in them. Justice therefore imposes an active duty: to disaggregate performance across groups, to ask who bears the cost of the system's errors, and to refuse the comfort of an aggregate metric that hides concentrated harm.

Explicability: the enabling principle

Explicability is the newest of the five and, in a sense, the one that makes the others enforceable. We cannot assess beneficence, detect harm, protect autonomy, or verify justice in a system we cannot understand. Explicability has two faces: *intelligibility*, the capacity to understand how a system works and why it produced a given output; and *accountability*, the capacity to identify who is answerable

for it. Neither requires that every model be a glass box—some opacity is the price of capability—but both require that the opacity be bounded, that consequential decisions be explainable to those they affect, and that responsibility never dissolve into "the algorithm did it."

These five principles recur throughout the guidelines as the standard against which specific practices are justified. They are not a checklist to be satisfied and forgotten; they are the questions a responsible team keeps asking, in different forms, at every stage of the work.

SECTION III

Ethics Across the Development Lifecycle

Ethical obligations are not a phase; they are a thread that runs the length of a system's life. The single most common structural error in responsible-AI programs is to treat ethics as a checkpoint—a review conducted once, usually late, after the expensive decisions have already been made. By the time a model is trained and a product is scoped, the questions that most determine its ethical character have been answered by default. The remedy is to distribute ethical attention across the whole lifecycle, so that each stage carries its own duties, its own owner, and its own artifact.

We organize these duties around six stages. The stages are sequential but the lifecycle is a loop: what is learned in operation feeds back into reframing, redesign, or retirement. Each stage below has a characteristic ethical question, a set of concrete obligations, and a record that documents how they were met—the artifact that turns a private judgment into an inspectable decision.

ETHICAL OBLIGATIONS ACROSS THE SIX-STAGE LIFECYCLE

Stage	Governing question	Key ethical obligation	Artifact
1. Framing	Should this be built, and for whose benefit?	Justify purpose; consider non-AI alternatives; identify affected groups	Purpose & impact memo
2. Data	Where did the data come from, and at what cost to whom?	Verify consent, provenance, representativeness, and minimization	Data statement / datasheet
3. Design & build	What values are we encoding into the system?	Make value-laden choices explicit; design for the five principles	Design & model card draft
4. Evaluation	Does it work fairly, safely, and as claimed?	Test disaggregated performance, robustness, and foreseeable misuse	Evaluation & harm-analysis report
5. Deployment	Is release justified, and under what conditions?	Weigh benefit against residual risk; set release and use limits	Release decision record
6. Monitoring & retirement	Is it still ethical as the world changes?	Watch for drift, misuse, and scope creep; retire responsibly	Monitoring log & sunset plan

Stage 1 – Framing: the cheapest ethical intervention

The most consequential ethical decision is often the first one: whether to build the system at all, and in service of whom. A direct interrogation at framing—what decision are we improving, who benefits, who could be harmed, and would a simpler intervention serve better?—dissolves many ill-conceived

projects before they consume resources or accrue momentum. Framing is also where the affected population is first identified, which is a precondition for engaging it (Section V) and for evaluating the system against it later. Scope defined carelessly here metastasizes into ethical trouble downstream.

Stage 2 — Data: inherited ethics

A model inherits the ethics of its data along with its statistics. At this stage a team establishes where the data originated, whether the people in it consented to this use, whom it represents and whom it omits, and what historical decisions—possibly unjust ones—are encoded in the labels. Data ethics is substantial enough to warrant its own treatment in Section IV; the lifecycle point is that these questions must be asked while the corpus is being assembled, not reconstructed defensively after a model trained on it has shipped.

Stage 3 — Design and build: values in the architecture

Every design choice encodes a value. The objective a model optimizes, the threshold at which it acts, the errors it is tuned to avoid, the interface through which a person meets it—each embeds a judgment about what matters and what may be sacrificed. The ethical discipline at this stage is to make those judgments explicit rather than to let them accrete as unexamined defaults. A team that can state what its system is optimizing, and what it is thereby willing to trade away, is practicing ethics; a team that cannot has delegated its values to whatever the tooling made easy.

A RECURRING PATTERN

Ethical failures are usually failures of a question not asked, at a stage where asking would have been cheap. "Whom does this leave out?" costs nothing during data collection and a great deal after launch. The lifecycle discipline exists to route each question to the stage where it is still inexpensive to answer honestly.

Stage 4 — Evaluation: testing for harm, not just accuracy

Evaluation in an ethical program tests more than whether the system performs. It tests whether it performs *fairly*—across the groups the framing stage identified—and whether it performs *safely* under the adversarial and out-of-distribution conditions it will actually meet. Crucially, it also tests for foreseeable misuse: not only "does it do what we intend?" but "what else will it do, and what will people do with it?" This harm analysis is where the dual-use questions of Section VII first become concrete, and its findings feed directly into the release decision that follows.

Stage 5 — Deployment: a decision, not an inevitability

Release is a choice that a named person makes and records, weighing the demonstrated benefit against the residual risk that evaluation surfaced. The decision is rarely binary. Between full release and no release lie staged rollouts, restricted access, use-based licensing, and capability limits—a

spectrum of conditions under which a system may be deployed responsibly. The release record documents not merely that release was approved but why, on what evidence, and under what constraints, so that the decision can be revisited if the evidence changes.

Stage 6 – Monitoring and retirement: ethics has no expiry

A system that was ethical at launch can drift into harm as the world moves around it—populations change, uses creep beyond the intended scope, and the assumptions of design quietly expire. Ongoing monitoring for performance drift, emergent misuse, and scope creep is therefore an ethical obligation, not merely an operational one. And when a system reaches the end of its useful life, retirement is itself an ethical act: the data it collected must be handled with care, the people who depended on its decisions must be considered, and the record of what it did must be preserved for those who may need to contest it after it is gone.

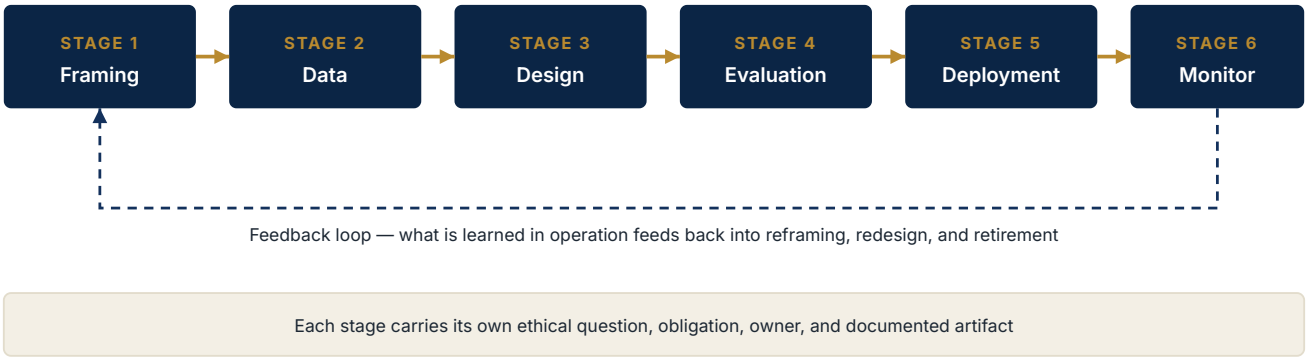

Figure 1. Ethical attention distributed across the development lifecycle. Ethics is a thread through every stage, closed into a loop by monitoring that feeds lessons back—not a single gate applied once at the end.

SECTION IV

Data Ethics — Consent, Provenance & Dignity

A model is a compression of its data. Whatever was taken without consent, whoever was omitted, whatever dignity was disregarded in the gathering—all of it is carried forward into the system and, through the system, into the lives it touches.

Data ethics is frequently the weakest link in an otherwise conscientious program, because data is acquired early, at speed, and under commercial pressure to gather as much as possible. Yet the ethical character of a system is substantially fixed at the moment its corpus is assembled. Four commitments organize responsible data practice: consent, provenance, minimization, and dignity. Each is a question a team must answer before, not after, a model learns from what it has collected.

Consent: the eroded foundation

Consent is the principle most strained by modern data practice. The web-scale corpora on which contemporary models are trained were, for the most part, not collected with the knowledge or agreement of the people whose words, images, and behavior they contain. "Notice and consent" has been stretched to a breaking point—buried in terms no one reads, invoked to justify uses no one anticipated. Ethical practice cannot restore a lost ideal, but it can insist on gradations: explicit, informed consent for sensitive and identifiable data; respect for the context in which data was originally shared; attention to whether a use would surprise or dismay the people concerned; and honesty about the cases where genuine consent was not, and could not have been, obtained.

Provenance: knowing what you are training on

Provenance is the discipline of knowing where data came from and being able to account for it. A team that cannot say what is in its training set cannot assess its representativeness, cannot respond when a source turns out to be tainted, and cannot honor a later request to remove someone's information. Provenance is not merely defensive record-keeping; it is a precondition for every other data-ethics judgment. The structured documentation practices the field has developed—datasheets for datasets, data statements—exist precisely to make provenance legible and to force the questions that ad-hoc collection skips.

DATA-ETHICS OBLIGATIONS BY SENSITIVITY OF THE DATA

Obligation	Public / low-sensitivity	Personal data	Sensitive / special-category
Consent basis documented	Source terms respected	Required	Required, explicit & informed
Provenance record	Source noted	Full datasheet	Full datasheet + review
Minimization applied	Recommended	Required	Required, strict necessity test
Representativeness analysis	Optional	Required	Required on affected groups
Removal / redress mechanism	Best effort	Required	Required, timely
Ethical review of collection	—	Recommended	Required

Minimization: the discipline of less

The instinct of data-driven development is to collect everything, on the theory that more data can only help. Ethics inverts the presumption: collect the minimum necessary for the stated purpose, and justify each additional category against it. Minimization reduces the harm of a breach, narrows the scope for mission creep, and forces a salutary clarity about what the system actually needs. Data not collected cannot be leaked, subpoenaed, repurposed, or used to profile the person who supplied it. In data ethics, restraint is a feature.

THE DIGNITY TEST

Behind every row of a dataset is a person who did not necessarily imagine being reduced to training signal. Before using data about people, ask whether the use honors their dignity—whether you would be comfortable explaining it to them directly. If the honest answer is that the use depends on their never finding out, that is itself the ethical finding.

Dignity: the person behind the data point

Consent, provenance, and minimization are procedural; dignity is the substantive value they serve. Data about people is not an inert resource like ore or bandwidth—it is a representation of human beings, often gathered at their most vulnerable, and the way it is handled either respects or diminishes them. Dignity resists full proceduralization, which is exactly why it matters: it is the check that remains when a use is technically permitted and formally consented to but still wrong. A dataset of images of people in crisis, a corpus of intimate conversations, a record of someone's medical or financial distress—each demands a care that no checkbox can capture. The dignity test asks not "are we allowed?" but "are we treating these people as ends rather than merely as means?"

> *Behind every data point is a person who never imagined being a data point.*
>
> — The discipline that data ethics is meant to restore

The practical upshot is that data ethics cannot be delegated wholesale to legal compliance. Compliance establishes the floor of what is permitted; ethics asks what is right, which is frequently a higher and more demanding standard. A team that treats "we had the legal right to use it" as the end of the inquiry has answered a lawyer's question, not an ethicist's.

SECTION V

Stakeholder Engagement & Participatory Design

The people most affected by an AI system are usually the last to be consulted about it, if they are consulted at all. A tool that screens tenants is designed without tenants; a system that allocates public benefits is built without the recipients; a content-moderation model shapes the speech of millions who never had a say in its rules. This asymmetry is not merely unfair—it is a practical defect, because the affected possess knowledge the builders lack, and their absence from the design table is a reliable predictor of the harms that follow.

Stakeholder engagement is the practice of correcting this asymmetry: of treating the people a system will touch as sources of design insight and as parties with a legitimate claim to be heard, rather than as passive recipients of whatever gets built. Done well, it is not a public-relations exercise appended after the fact but a design input that shapes the system while shaping is still possible. Done badly—consultation staged after the architecture is frozen—it is worse than nothing, because it lends the appearance of legitimacy to a process that ignored the very people it performed listening to.

Who counts as a stakeholder

The stakeholder map is wider than the customer list. It includes the direct users of a system, but also the people who are subject to its decisions without ever choosing to use it—the loan applicant, the flagged traveler, the moderated poster. It includes the workers whose jobs the system reshapes, the communities on whom its errors concentrate, and the domain experts whose tacit knowledge the model attempts to encode. A rigorous engagement begins by drawing this map deliberately, with particular attention to those who are affected but not represented—because the groups least able to demand a seat are frequently those most exposed to harm.

A LADDER OF STAKEHOLDER ENGAGEMENT — FROM PERFORMANCE TO PARTNERSHIP

Level	What happens	What the stakeholder actually gets
1 · Inform	The team tells stakeholders what it has decided.	Notice after the fact; no influence on the outcome.
2 · Consult	The team solicits opinions, then decides alone.	A voice, but no assurance it changed anything.
3 · Involve	Stakeholder input demonstrably shapes design choices.	Documented influence on specific decisions.
4 · Collaborate	Stakeholders share decision-making on key questions.	Partnership; shared authority over trade-offs.
5 · Co-design	Affected people help define the problem and the solution.	Ownership; the system reflects their priorities.

The ladder is a diagnostic, not a mandate that every project climb to the top. Low-stakes tools may legitimately sit at "inform" or "consult." But a system that makes consequential decisions about people's lives has an obligation to reach at least "involve"—to be able to point to specific design choices that the affected changed. The honest question for any project is not whether it engages stakeholders but which rung it reaches, and whether that rung is proportionate to what is at stake.

Participatory design in practice

Participation is easier to endorse than to do well, and it fails in characteristic ways. It fails when it is *tokenistic*—a single community member on a panel of thirty, present to be counted rather than heard. It fails when it is *extractive*—taking people's time and knowledge without compensation, feedback, or any share in the result. And it fails when it is *late*—staged after the decisions that mattered were already made. Genuine participation compensates people for their contribution, engages them early enough to influence architecture rather than only cosmetics, closes the loop by reporting back what their input changed, and shares power over at least some real decisions.

A HARD-WON LESSON

"Nothing about us without us"—the demand that emerged from disability-rights advocacy—is a reliable guide for AI. If a system will make decisions about a group, that group belongs in the room where the decisions about the system are made. Their absence is not a neutral omission; it is a design choice with predictable consequences.

When engagement changes the answer

The strongest sign that engagement is genuine rather than performed is that it sometimes changes the outcome—that the team builds something different, narrows a scope, adds a safeguard, or abandons a project because the people it consulted told it something it did not want to hear. Engagement that never alters a decision is not engagement; it is choreography. Teams should expect, and welcome, the friction of being told that their idea is unwelcome, that their convenient data source is a source of harm, or that the problem they set out to solve is not the problem the affected community actually has. That friction is the value. It is cheaper to encounter it in a design session than in a deployment.

SECTION VI

Ethical Review — An IRB for AI

Medicine confronted the problem of well-intentioned researchers causing harm decades ago, and answered it with a durable institution: independent ethical review before consequential work proceeds. AI needs its own version—adapted, proportionate, and enforceable.

The institutional review board is one of the more successful ethical inventions of the twentieth century. Born from the failures of human-subjects research, it establishes a simple, powerful discipline: before certain research proceeds, an independent body competent to judge its ethics reviews the plan and can require changes or refuse approval. The IRB is not perfect, and importing it uncritically into AI would be a mistake—much AI development is not human-subjects research in the classical sense, and a board that reviews everything will become a rubber stamp. But the core idea—independent, competent, documented review proportionate to the stakes, conducted before the consequential decisions are locked—transfers directly and is badly needed.

Triggers: what requires review

The first design question is not who sits on the board but what brings a project to it. Review is a scarce resource; spending it on everything wastes it. A workable trigger set routes to review the projects where the stakes justify the scrutiny, and lets the rest proceed under lighter-weight guidance. The triggers below are illustrative; each organization should calibrate them to its domain, but the principle is constant: the trigger is the stakes, not the technology.

ETHICAL-REVIEW TRIGGERS AND THE TIER OF REVIEW THEY INVOKE

If the system...	Example	Review tier
Makes or heavily steers decisions about people	Benefits eligibility, hiring, credit, risk scoring	Tier 3 — Full board
Uses sensitive or special-category data	Health, biometrics, children, protected traits	Tier 3 — Full board
Could plausibly be misused for serious harm	Surveillance, synthetic media, capability uplift	Tier 3 — Full board
Affects a vulnerable population	Patients, minors, marginalized communities	Tier 2 — Expedited
Involves consequential automation with a human in the loop	Triage prioritization, investigative leads	Tier 2 — Expedited
Is a low-stakes productivity or internal tool	Drafting aids, internal search, summarization	Tier 1 — Self-assessment

The board: independence and competence

An ethical review board is only as good as its independence and its competence, and both must be engineered. Independence means the board does not report to the people whose projects it reviews and is not rewarded for approving them; a board that answers to the shipping deadline will find reasons to approve. Competence means the board can actually understand what it reviews, which requires a deliberate mix: technical expertise to grasp what a system does, domain expertise in the affected field, ethics and legal knowledge, and—critically—representation from or genuine connection to the communities the work affects. A board composed entirely of the organization's own senior engineers is neither independent nor representative, however competent.

A **Independence.**
The board's authority and standing do not depend on the goodwill of the teams it reviews. Its members are protected when they say no, and saying no is a real, recorded possibility.

B **Competence.**
The board's composition matches the work: technical, domain, ethical, and legal expertise, plus a genuine connection to the affected communities. It can call on outside experts when it cannot judge a question itself.

C **Authority.**
The board can require changes and can withhold approval, and its determinations bind. A review body whose conclusions are advisory to the people it reviews is a suggestion box, not a control.

Documentation: the determination that leaves a trace

The output of review is a written determination: what was proposed, what concerns were raised, what conditions were imposed, and why approval was granted, withheld, or made contingent. Documentation serves three ends. It creates *accountability*, fixing who decided what and on what basis. It enables *learning*, so that an organization can see patterns across its own decisions and improve its guidance. And it provides *evidence* of diligence to regulators, courts, and the public. A review that leaves no record is, for every practical purpose, a review that did not happen.

PROPORTIONALITY, AGAIN

The failure mode of ethical review is not excessive rigor on the dangerous cases; it is uniform rigor on all cases, which drives teams to route around the board and drains its legitimacy. Self-assessment for the trivial, expedited review for the consequential, full board for the grave—this is what keeps serious review affordable enough to remain serious.

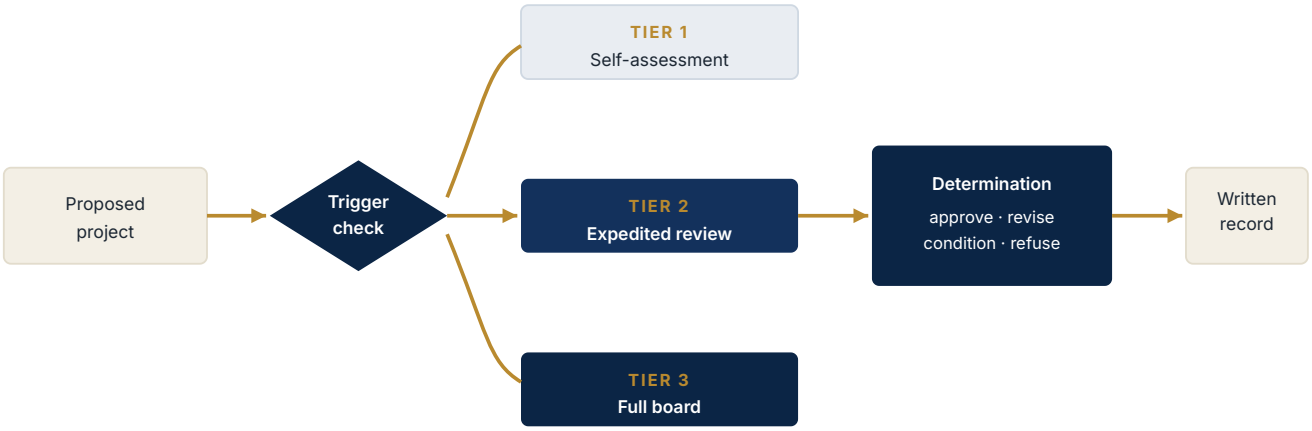

Figure 2. The ethical-review workflow. A trigger check routes each project to a proportionate tier; every consequential review ends in a documented determination and a written record that survives the project.

SECTION VII

Dual-Use & Downstream Harm

A system built to do good can be turned to do harm, and often the same capability serves both. A model that detects disease in medical images can be repurposed to identify people from them; a tool that generates realistic speech can restore a voice to someone who lost theirs or clone a voice to commit fraud; a system that finds vulnerabilities to patch them can find them to exploit them. This is the dual-use problem, and it is not a niche concern for weapons research—it is a routine property of powerful, general technology, and it demands foresight that most development processes do not supply.

The ethical obligation here extends the principle of non-maleficence into a harder register. It is not enough to ask whether a system works as intended, or even whether it fails in foreseeable ways. One must ask how it will be used by people whose intentions differ from the builder's—the motivated misuser, the opportunistic repurposer, the adversary who reads the same paper the defenders do. This foresight is uncomfortable, because it requires imagining one's creation in hostile hands, but it is precisely the discomfort that makes it valuable.

Anticipating misuse: the abuse case

Engineering has long used "use cases" to describe intended behavior. Ethical AI development needs their shadow: the *abuse case*, a deliberate exercise in imagining how a system will be misused. The method is structured pessimism. For a given capability, ask who would want to misuse it and to what end; how they would go about it; what the system makes easier for them than it was before; and whether the marginal uplift the system provides to a bad actor is offset by the good it does for legitimate users. This last calculus—the balance of benefit to defenders against uplift to attackers—is often the crux of a release decision, and it deserves to be made explicitly rather than assumed away.

A FRAMEWORK FOR REASONING ABOUT DUAL-USE RISK

Question	What it probes	Illustrative consideration
Who would misuse it?	The threat model	Fraudsters, harassers, authoritarian states, careless amplifiers
How much does it help them?	Marginal uplift	Does it enable new harm or merely restate what is already easy?
How much does it help the many?	Legitimate benefit	Is the good broad and real, or thin and speculative?
Can the misuse be mitigated?	Available safeguards	Access controls, use restrictions, provenance signals, monitoring
Is the harm reversible?	Durability of damage	Reputational and safety harms may not be undoable after the fact

Release as a spectrum

The dual-use question is frequently framed as a binary—release or withhold—but that framing forecloses the most useful options. Between open release and total non-publication lies a spectrum of graduated choices: staged release that begins with restricted access and widens as risks are understood; structured access through interfaces that permit use while constraining misuse; use-based licensing that permits benign applications and prohibits harmful ones; capability limits designed into the system; and release of a model without the training details that would let it be reconstructed for other ends. The right point on this spectrum depends on the balance of uplift and benefit, and choosing it deliberately is itself an ethical act.

THE FORESIGHT ASYMMETRY

Builders imagine their systems being used well, because that is why they built them. Misusers imagine them being used badly, because that is what they want. The abuse case is a deliberate correction for this asymmetry—an appointed moment to think like the adversary before the adversary has to teach you.

Downstream harm and the limits of foresight

Not all harm is deliberate misuse. Some is the emergent, systemic consequence of a system working exactly as designed: a recommendation engine that optimizes engagement and, in doing so, rewards outrage; a labor-automating tool that delivers efficiency to an employer and precarity to a workforce; a pattern of individually reasonable decisions that aggregates into a discriminatory whole. These downstream harms are harder to anticipate because they arise from interactions—between the

system and a market, a society, a population—that no single team fully controls. Foresight has limits, and honesty about those limits is itself an ethical stance.

What foresight's limits imply is not resignation but humility expressed as process. Because we cannot foresee everything, we build the capacity to respond: monitoring that watches for the harms we did not predict, channels through which affected people can report what we missed, and a genuine willingness to change or withdraw a system when the evidence arrives. The measure of a responsible builder is not the impossible standard of perfect prediction but the achievable one of vigilant response—of noticing emergent harm early and acting on it rather than defending the original decision.

> *The question is not only "does it work?" but "what will people do with it when it does?"*
>
> — The discipline of dual-use foresight

SECTION VIII

Operationalizing Ethics & Escaping "Ethics Washing"

An ethics program lives or dies not on the quality of its principles but on the structures that make those principles bite: the roles that own them, the incentives that reward them, and the metrics that reveal whether they are working. This final section is about making ethics real.

Every preceding section has translated a principle into a practice. This one confronts the reason practices fail even when they are well designed: they are bolted onto an organization whose roles, incentives, and metrics all point the other way. An ethics process that asks people to slow down in an organization that rewards only speed will lose every time the two conflict. Operationalizing ethics means aligning the machinery of the organization—who is responsible, what gets rewarded, what gets measured—with the commitments the ethics statement makes.

Roles: giving ethics an owner

Responsibility that belongs to everyone belongs to no one. Ethical practice requires named owners at several levels: individual practitioners accountable for the ethics of their own work; a review body (Section VI) with independent authority over consequential decisions; and senior leadership that owns the program's resourcing and its standing when it collides with business pressure. The most common structural failure is to appoint an ethics function with responsibility but no authority—a team that can raise concerns but cannot stop anything, whose objections are heard and then overruled by the people with real power. Responsibility without authority is not accountability; it is a fuse designed to blow quietly.

1. **Practitioners own their work.**
   Every builder is accountable for the ethical questions their own decisions raise, and has the training and the standing to raise them without penalty.

2. **Review holds real authority.**
   The review body can require changes and withhold approval, and its determinations are not routinely overruled by the teams it reviews.

3 **Leadership owns the incentives.**
Senior leaders are accountable for whether the incentive structure rewards or punishes the ethical behavior the organization professes to want.

4 **Someone can say stop.**
At least one person has the authority to halt a launch on ethical grounds and is protected for using it in good faith. An ethics program with no off-switch is advisory.

Incentives: the true statement of values

An organization's real ethics are written in its incentives, not its principles document. If people are promoted for shipping and never for the launch they responsibly delayed, the organization has declared that speed matters and ethics does not, whatever its website says. Aligning incentives is the hardest and most important part of operationalizing ethics, because it requires the organization to pay a real price—sometimes a slower launch, a foregone data source, a declined project—and to reward the people who impose that price for good reason. Until raising an ethical concern is career-neutral at worst and career-enhancing at best, the concerns will go unraised, and the process built to surface them will quietly starve.

Metrics: measuring whether it works

What is not measured is not managed, and ethics is unusually easy to leave unmeasured because its successes are non-events—the harm that did not occur, the launch that was responsibly reshaped. Useful metrics are therefore mostly about process and its consequences rather than about outcomes alone: how many projects triggered review and what the reviews changed; how often concerns were raised and how they were resolved; how disaggregated performance looks across the groups a system serves; how quickly reported harms are addressed. These measures are imperfect and gameable, and they must never become targets that substitute for judgment. But their absence is telling: an organization that measures its velocity in exquisite detail and its ethical practice not at all has revealed which one it takes seriously.

4
levels of ownership, from practitioner to leadership, each accountable

3
alignment levers—roles, incentives, and metrics—that make principles bite

1
non-negotiable: someone with the authority to say stop

Escaping ethics washing

Ethics washing—the performance of ethical concern as a substitute for its practice—is the characteristic failure of this field, and it is corrosive precisely because it is cheaper than the real thing and superficially indistinguishable from it. A glossy principles document, a well-publicized advisory board, a conference talk on responsible AI: each can be genuine or each can be theater, and from the outside they look the same. What distinguishes practice from performance is not visible in the statement; it is visible in the record. Does the review board ever say no? Has an incentive ever been changed to reward caution? Can the organization point to a launch it delayed, a dataset it refused, a project it stopped? Ethics that has never cost anything has never done anything.

THE TEST THAT CANNOT BE GAMED

Ask an organization for its principles and you learn what it wants to be seen believing. Ask it for the last three times ethics changed a decision—and for the record of who decided, on what evidence, at what cost—and you learn whether it believes anything at all. The first question invites a brochure; the second invites the truth.

A concluding word to builders

The through-line of these guidelines is that ethics is a practice, not a posture—something a team does, repeatedly and at cost, rather than something it declares. The principles in Section II are durable, but they are inert until a process gives them force: obligations attached to lifecycle stages, care exercised over data, the affected brought into design, consequential work put to independent review, misuse anticipated before it arrives, and an organization arranged so that doing right is rewarded rather than punished. None of this makes a team slower where slowness is not warranted; it makes a team trustworthy where trust is earned. The goal is not the appearance of responsibility. It is the reality of it—and the difference, in the end, is borne by the people these systems touch.

CLOSING NOTE

The aim of an ethics program is not to prevent teams from building. It is to let them build the many benign things freely, precisely because they have the discipline to build the consequential things carefully. Ethics practiced well is not the enemy of good engineering. It is what good engineering, extended to its human consequences, becomes.

APPENDIX

Notes & Further Reading

These guidelines synthesize established research ethics with the contemporary literature on AI ethics, documentation, and governance. The sources below are entry points into that literature; they are indicative rather than exhaustive.

1. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research.* The foundational statement of respect for persons, beneficence, and justice that underlies modern ethical review.
2. Organisation for Economic Co-operation and Development. *OECD AI Principles.* Intergovernmental principles for trustworthy AI, including human-centered values, transparency, and accountability.
3. IEEE. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems.* A comprehensive framework linking human rights and well-being to system design.
4. Université de Montréal. *Montréal Declaration for a Responsible Development of Artificial Intelligence.* A participatory declaration of principles developed through public deliberation.
5. Mitchell, M., et al. *Model Cards for Model Reporting.* On structured documentation of a model's intended use and performance across conditions and groups.
6. Gebru, T., et al. *Datasheets for Datasets.* On documenting data provenance, composition, collection, and intended use.
7. Bender, E. M., & Friedman, B. *Data Statements for Natural Language Processing.* On documenting the characteristics and limits of language data.
8. Floridi, L., & Cowls, J. *A Unified Framework of Five Principles for AI in Society.* On the convergence of AI principles onto beneficence, non-maleficence, autonomy, justice, and explicability.
9. Beauchamp, T. L., & Childress, J. F. *Principles of Biomedical Ethics.* The canonical account of the four principles adapted here for AI.
10. Metcalf, J., & Crawford, K. *Where Are Human Subjects in Big Data Research? The Emerging Ethics Divide.* On the gap between research-ethics review and data-driven development.
11. Selbst, A. D., et al. *Fairness and Abstraction in Sociotechnical Systems.* On why fairness and ethics cannot be resolved at the level of the model alone.
12. Bietti, E. *From Ethics Washing to Ethics Bashing.* On the performance of ethics as a substitute for its practice, and how to tell the difference.