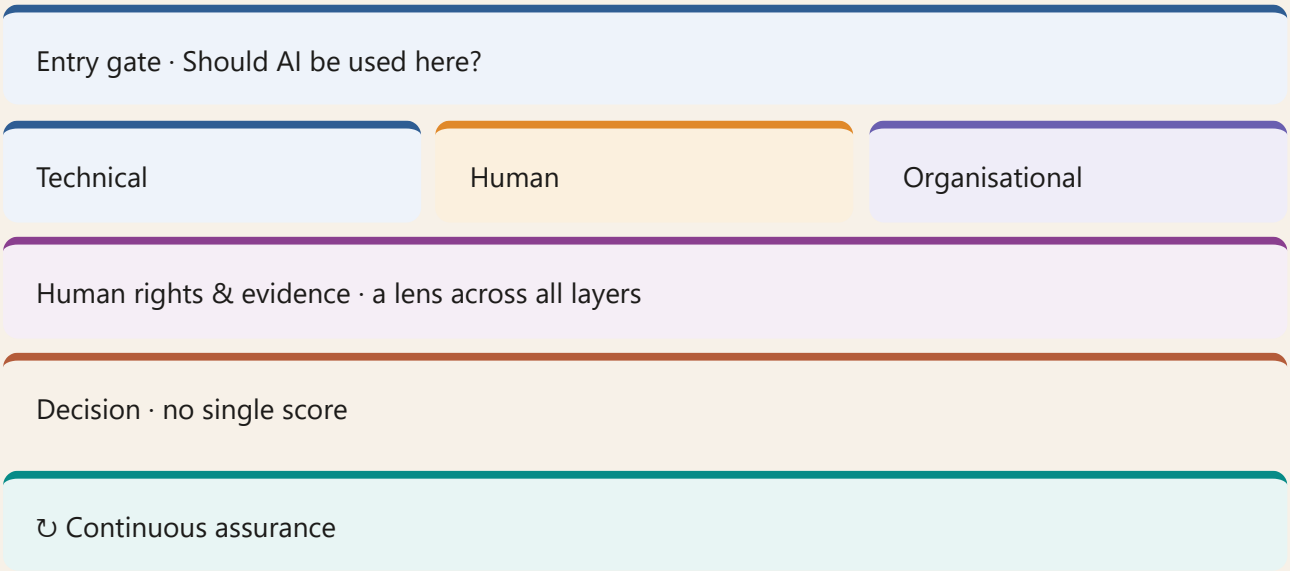
FRAMEWORK OVERVIEW

Responsible AI & Human Agency Framework

An evidence-based method for assessing not only whether AI works responsibly, but what its use does to human agency, capability, dignity and accountability.

In development · Working draft v0.5 · October 2026



Daphne Iris van Vliet
Independent research

daphneiris.com/writing/responsible-ai/framework

Not a certification scheme · not legal advice

READ THIS FIRST

What this document is — and is not

It is

- An overview of a framework I am developing in the open.
- A synthesis of ten research studies, with a method for turning them into an assessment.
- Independent and vendor-neutral.

It is not

- A validated method. Domains, indicators and decision rules are proposals.
- A certification, audit or compliance scheme — and not legal advice.
- Endorsed by any employer, supplier, institution or standards body it references.

Please do not use v0.x versions to certify, approve or reject an AI system. The framework will change as it is verified and piloted.

THE CENTRAL QUESTION

Technically responsible — and humanly sound?

How do we establish whether an AI system is technically responsible while preserving human autonomy, capability, dignity and accountability?

An AI system can meet its performance targets and its legal obligations while still encouraging excessive reliance, eroding professional judgment, excluding particular groups or making responsibility harder to locate. So the framework assesses two things at the same time:

> the quality of the technology — and the quality of the human situation it creates, for the people who use it, the people affected by it and the organisation deploying it.

Three convictions shape the method:

- **The fact that AI can perform a task does not establish that it should.** Assessment starts by asking whether AI belongs here at all.
- **Claims need evidence.** A policy saying that humans stay in control is an assertion, not a demonstration.
- **There is no single score.** Serious failure in one area cannot be averaged away by strength in another.

Read online: The framework, online

ARCHITECTURE

A gate, three layers, a lens and a loop

Entry gate · Should AI be used here?

Proportionality and purpose integrity: is AI necessary, suitable and the least intrusive option — and what must remain human?

↓

1 · Technical

Can the technology be trusted for this use?

2 · Human

What does it do to people and their capabilities?

3 · Organisational

Can the organisation demonstrate responsible use?

Human rights & evidence · a lens across all three layers

Are the rights at stake actually protected in practice — not only on paper?

↓

Decision rules

Red lines block · weakest link, not average · uncertainty raises the bar · the human layer cannot be waived · every decision expires

↓

↻ Continuous assurance

Monitor · challenge · learn → reassess. New evidence can send a use back to the entry gate.

The gate comes first because the layers can only say *how* to use AI responsibly. Only the gate can say *whether* to. A use that fails it is not assessed further: the outcome is **do not use** or **human-primary**.

Human rights are deliberately not a fourth layer. Privacy, non-discrimination, autonomy and remedy each run through the technical, human and organisational layers at once, so they are assessed as a cross-cutting lens.

Read online: Architecture · Human rights & evidence

UNIT OF ASSESSMENT

A use, not a model

The same model can be reasonable for summarising internal meetings, high-risk in recruitment and inappropriate for automated welfare decisions. So the framework never assesses "model X". It assesses a specific **AI use**, defined by six elements:

Purpose
What it is for — and not for.

System
Model, version, configuration, data, sources, tools, supplier.

Workflow
Where AI sits, and where authority sits.

People
Users, people affected, people who carry the work or risk.

Organisation
Owner, incentives, oversight, remedy.

Context
Sector, law, stakes, reversibility.

A material change to any element is a reason to reassess.

ENTRY GATE

The AI Use Proportionality Test

1. **Purpose** — what legitimate human or organisational purpose is served?
2. **Value** — is there credible evidence AI improves on the best realistic alternative?
3. **Necessity** — could a less intrusive option deliver comparable value?
4. **Human impact** — what happens to rights, autonomy, capability, dignity and equality?
5. **Control** — are errors detectable, contestable and reversible under real human authority?
6. **Distribution** — who gains, and who carries the risk, work and consequences?
7. **Purpose integrity** — does the activity keep the human good it exists to provide?

Efficiency alone is not enough. Uncertainty raises the burden of proof. The human-only baseline must be tested too.

Read online: Study 01 · When Should AI Not Be Used?

PURPOSE INTEGRITY

What disappears when AI succeeds?

Some activities produce an outcome *and* a human good that exists only because of how the outcome is produced.

ACTIVITY	PRODUCES…	…AND A GOOD CREATED BY THE PROCESS
Care	Treatment	Being attended to
Education	Answers	Learning, formation and authorship
Law	Decisions	Reason-giving, procedural justice, legitimate authority
Management	Allocation	Recognition, coaching and responsibility
Public consultation	Information	Democratic participation

Automation is **purpose-undermining** when it improves a measurable proxy — speed, throughput, consistency — while hollowing out that underlying good. Where the human core cannot be kept without eliminating the claimed benefit, the outcome should normally be human-primary or do not use.

Human-primary is a real category

AI may retrieve, compare, summarise, translate, flag and structure. A human interprets, deliberates, decides, justifies and takes responsibility. The question is not whether a human is "in the loop", but where authority sits.

FIVE LEGITIMATE OUTCOMES

Do not use
Unlawful or illegitimate purpose, uncontrollable harm, impossible oversight or no remedy.

Human-primary · AI-assist
AI adds value, but judgment and responsibility stay human.

Restricted pilot
Plausible benefit, insufficient evidence; bounded, with stop rules.

Strong safeguards
High-stakes use with demonstrated value, oversight, remedy and rollback.

Proportionate use
Low or controlled risk, net value, reversible effects, fair distribution.

Responsible governance is not a one-way ratchet toward more automation: reassessment can move a use in either direction.

Read online: Purpose integrity · Human-primary

WHAT IS ASSESSED

Eighteen domains in three layers

Two entry-gate tests come first. Then each layer has six assessment domains, each with a core question and at least one observable indicator. The human layer has equal structural status to the technical and organisational layers — it is not an appendix.

Entry-gate test · Proportionality

Is AI necessary, suitable and the least intrusive option for a legitimate purpose?

Entry-gate test · Purpose integrity

Does the activity keep the human good it exists to provide?

1 · Technical

Validity & accuracy
Does it work for this task, population and setting?

Robustness & reliability
Does performance hold under change and stress?

Fairness & subgroups
Who does it serve worse?

Uncertainty & abstention
Can it signal when not to rely on it?

Security & privacy
Are access, data and actions proportionate and protected?

Traceability
Can an output or action be reconstructed?

2 · Human

Retained capability
Can people still do the work, and challenge outputs, without AI?

Appropriate reliance
Do people accept correct outputs and reject wrong ones?

Meaningful oversight
Do reviewers have the information, time, competence and authority to disagree?

Autonomy & choice
Can people refuse or reach a human without penalty?

Dignity & relational integrity
Does the design misrepresent competence, agency or care?

Human impact & distribution
Where does saved time go, and who carries the cost?

3 · Organisational

Accountability
Is there a named owner who can answer, decide and stop?

Incentives & culture
Do targets reward responsible outcomes, or only speed and adoption?

Participation & inclusion
Did affected people have influence before the decision?

Legal compliance
Are legal duties met — as a floor, not the finish line?

Independent challenge
Can someone without a stake question it?

Incident response & remedy
Can harm be reported, corrected and remedied?

Measuring people without surveilling them. Human indicators are collected at team or cohort level, with consent, minimal data and a clear purpose — never turned into individual performance telemetry.

Read online: Indicators for every domain · All domains with their studies

EVIDENCE

Evidence classes, not a ladder

CLASS	QUESTION IT ANSWERS
E0 · Assertion	What is claimed? A policy, supplier or team says it is true.
E1 · Controlled evidence	What happens under defined test conditions?
E2 · Production evidence	What happens in actual use?
E3 · Independent challenge	Does the claim survive scrutiny from outside the deploying team?
E4 · Human corroboration	Does the protection hold from the perspective of the people affected?

The classes are complementary. E4 is not automatically stronger than E3: an adversarial test is far stronger evidence of data leakage than interviews, while no technical audit can show whether affected people can actually contest a decision. **Evidence is triangulated according to the claim.** E0 alone never supports a positive finding.

An evidence class describes the *kind* of evidence — not legal weight, maturity or compliance. A binding legal duty can be met with E0 evidence; a voluntary principle can call for E4 evidence.

Six evidence perspectives

PERSPECTIVE	QUESTION	EXAMPLE: AN APPEAL ROUTE
Structural	Are rules, authority and safeguards in place?	There is an appeal process.
Process	Are the safeguards being implemented?	Appeals are processed.
Behavioural*	Can and do people exercise the capability or right?	People can successfully appeal.
Outcome	What actually happens to people?	Wrong decisions are corrected.
Experience*	How do affected people experience it?	The process is understandable and usable.
Distributional*	Do burdens differ across groups?	Some groups abandon appeals far more often.

*Structural, process and outcome indicators follow the UN human rights indicator method (OHCHR, 2012). *Behavioural, experience and distributional perspectives are my proposed extension for AI.*

Read online: Evidence classes · Six evidence perspectives

FROM EVIDENCE TO DECISION

Findings without a score

Each domain receives one finding, recorded with its evidence classes:

Demonstrated

The claim is supported by evidence appropriate to the stakes.

Partly evidenced

Supported, but with material gaps or below the required evidence.

Not evidenced

No adequate evidence either way. In higher-stakes uses this constrains deployment.

Red flag

Evidence of serious harm, or a red line is crossed.

Five decision rules

1 **Red lines block.** A prohibited practice, illegitimate purpose or severe irreversible harm without remedy means do not use.

2 **Weakest link, not average.** Strength in one domain cannot offset a red flag in another.

3 **Uncertainty raises the bar.** Higher stakes, vulnerability and irreversibility demand stronger evidence.

4 **The human layer cannot be waived.** Strong technical findings with unassessed human domains are not ready for scale.

5 **Every decision expires.** Each names its conditions, owner and review date.

An evidence profile, not "72%"

Privacy	Demonstrated Necessity review, access logs, replicated leakage test.
Non-discrimination	Partly evidenced Too few cases for two materially affected groups.
Autonomy	Not evidenced An opt-out exists in policy; nothing shows it is usable.
Contestability	Red flag An appeal exists, but reviewers cannot change the outcome.

A hypothetical illustration. Nothing is averaged: strong privacy cannot cancel an ineffective remedy.

Read online: Decision rules · Findings per right

MEASURING HUMAN AGENCY

Declared control is not demonstrated agency

> Do not ask only whether people feel able to exercise agency. Test whether they can.

Perception | Knowledge | Capability | Behaviour | **Outcome**

CLAIM	WEAK EVIDENCE	STRONGER EVIDENCE
People understand the AI	They say they do	They identify its limits and when to escalate
Humans remain in control	Policy allows override	Seeded wrong recommendations are actually rejected
Expertise is kept	Training completed	Unaided performance holds after sustained use
Reliance is appropriate	A trust score	Correct advice accepted, wrong advice resisted
People can challenge decisions	An appeal link exists	Wrong decisions are found, challenged and corrected

Two people, two kinds of agency

The operator or decision-maker

Understand → judge → challenge → intervene → retain capability

The affected person

Notice → understand → add context → reach a human → contest → correct → obtain remedy

A low override rate may mean an excellent system — or passive reviewers. Zero complaints may mean zero harm — or that nobody knows how to complain. Measures only acquire meaning through triangulation. Six candidate dimensions and fourteen pilot indicators are set out online; they are being tested, not settled.

Read online: Measuring Human Agency in AI Systems

HUMAN RIGHTS & EVIDENCE

A metric is not a right

> A metric can provide evidence relevant to a human right. It cannot, by itself, prove that the right has been respected.

The Rights-to-Evidence method

Right · Protected human condition · Observable construct · Indicators · Evidence · Interpretation · Finding · **Decision**

Start from the right, not from whatever data is easiest to collect. Ten rights domains are examined: privacy, equality and non-discrimination, autonomy, dignity, accessibility, participation, contestability and remedy, procedural fairness, labour rights, and expression and information.

Rights in practice

Formally available · Findable · Understandable · Exercisable · Changes the process · Corrects the outcome · **Remedies the harm**

No detected disparity ≠ non-discrimination demonstrated

It may reflect small samples, the wrong groups, missing data or harm later in the workflow.

A measured disparity ≠ unlawful discrimination

It is evidence that needs investigation, legal interpretation and examination of justification.

> Rights-respecting measurement: the method used to measure a human right must itself respect human rights.

Collect only what the assurance question needs; aggregate; restrict access and retention; never publish small groups; never protect worker autonomy by monitoring every worker.

Read online: Measuring Human Rights in AI

RELATION TO LAW AND STANDARDS

Complement, not replace

The framework builds on the EU AI Act, the GDPR, the NIST AI Risk Management Framework 1.0, ISO/IEC AI standards (42001, 23894, 42005, 42006), the OECD AI Principles and the UNESCO Recommendation on the Ethics of AI. All eighteen domains have been mapped against them, provision by provision.

> Existing instruments provide substantial coverage of technical risk, governance processes, fundamental rights and selected human impacts. Coverage is less systematic and less operational for retained human capability, appropriate reliance, relational integrity, practical autonomy, the distribution of gains and burdens, organisational incentives, and the influence affected people have over AI-use decisions.

Largely covered — reuse existing evidence

Validity, robustness, security and privacy, traceability, meaningful oversight where AI Act Article 14 applies, accountability as a role, incident response, lifecycle monitoring.

Where the framework adds most

Retained capability · appropriate reliance · uncertainty and abstention · autonomy and choice · dignity and relational integrity · incentives · participation with real influence.

Important limits

- **Coverage is not applicability.** Many AI Act duties apply only to high-risk systems, and the GDPR only to personal-data processing.
- **Versions matter.** NIST mappings are locked to AI RMF 1.0, which NIST states is being revised.
- **ISO clauses are pending.** ISO is mapped at standard level; clause-level mapping awaits verification against licensed text.
- **Not legal advice.** Most AI Act high-risk obligations apply from 2 December 2027 (Annex III) and 2 August 2028 (Annex I).

Read online: The full standards mapping

OVER TIME

Evidence expires

> Evidence supporting a responsible-use decision expires when the facts that justified it materially change.

Models change, suppliers update, people rely more, skills fade, appeal rates rise, benefits fall short. Continuous assurance asks whether a use should *still* be running — across all eighteen domains, not only technical performance.

Normal | Watch | Review | Intervene | **Suspend or withdraw**

Every material approval is time-limited, with a named owner who must actively re-authorise it. A programme whose evidence never changes a decision may be successful — or ceremonial.

WHO ASSESSES

A system cannot certify itself

ROLE	RESPONSIBILITY
Use owner	Provides evidence and is accountable — but does not sign off alone.
Independent assessor	Tests the evidence, sets findings, has stop and escalation rights, no stake in the deployment.
Users and frontline staff	Supply evidence on workload, reliance, capability and whether override is real.
Affected people	Shape success criteria and purpose integrity before the decision.
Worker representatives	Review workplace effects, surveillance risk and the distribution of gains.
Legal, privacy and security	Confirm the legal floor and the red lines.

Every assessment records conflicts of interest, including commercial relationships with the AI supplier.

Read online: Study 10 · Continuous assurance · Who assesses

APPLYING THE FRAMEWORK

Instruments and pilots

Organisational Questionnaire

Gathers claims, documents, experience and operational evidence through a 32-question rapid screen and nine stakeholder tracks.

Practitioner Workbook

Where an assessor tests claims, triangulates evidence, records uncertainty and red flags, and writes a decision record.

> Respondents provide evidence. Assessors make findings. Decision-makers authorise or refuse use.

Status: both instruments are a research prototype, previewed online and not yet released for use.

Three contrasting pilots

SETTING	TESTS MOST
Enterprise AI assistant	Retained capability, distribution of gains, incentives
AI-assisted recruitment	Fairness, participation, oversight, contestability
Customer-facing conversational AI	Relational integrity, appropriate reliance, autonomy

> A framework that never changes a decision has not been validated — it has been rehearsed.

Release path

v0.x · Research prototype

Now. Method in development; not for certification.

v0.9 · Public beta

Verified and piloted; released for structured testing.

v1.0 · Validated release

Revised after pilots, reliability testing and review.

Read online: Practitioner instruments · Pilot protocol

RESEARCH FOUNDATION

Ten studies behind the framework

01 When Should AI Not Be Used?
Entry gate · proportionality · purpose integrity

02 Can an AI System Be Legally Compliant Yet Still Irresponsible?
Governance

03 Whose Interests Are Represented in AI Design?
Inclusion and participation

04 When Does AI Assistance Become Human Dependency?
Human capability

05 Meaningful Human Oversight in AI
Human oversight

06 Who Is Accountable When Responsibility Is Distributed?
Accountability

07 Measuring the Human Cost of AI-Driven Efficiency
Human impact

08 Why Do Responsible AI Commitments Fail Under Pressure?
Incentives

09 What Makes a Relationship With AI Trustworthy?
Trust and appropriate reliance

10 How Do We Know AI Remains Responsible Over Time?
Continuous assurance

Read the studies, the standards mapping, the indicators and the measurement methods — with all sources — at **daphneiris.com/writing/responsible-ai**.

Independence

This framework is my own synthesis. It is not affiliated with, commissioned by or endorsed by any employer, supplier, institution or standards body, including those referenced.

How to cite

van Vliet, D. I. (2026). *Responsible AI & Human Agency Framework — Overview*. Working draft v0.5, in development. daphneiris.com/writing/responsible-ai/framework.

Version 0.5 · October 2026. The framework changes as evidence accumulates; check the online version history for the current version.