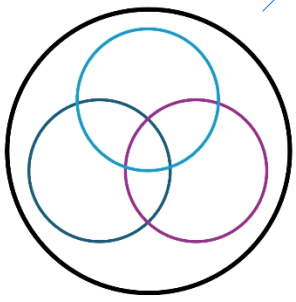
POTENTIALISM FRAMEWORK

Version 2.0



By Emad Sadeghipour & AI Systems
May 7, 2026

Table of Contents

CC BY–NC 4.0

CC BY–NC 4.0

CC BY–NC 4.0

CC BY–NC 4.0

CC BY–NC 4.0

CC BY–NC 4.0

CC BY–NC 4.0

CC BY–NC 4.0

CC BY–NC 4.0

CC BY–NC 4.0

CC BY–NC 4.0

CC BY–NC 4.0

CC BY–NC 4.0

CC BY–NC 4.0

CC BY–NC 4.0

CC BY–NC 4.0

CC BY–NC 4.0

0 Front matter

CC BY–NC 4.0

0.1 Abstract

- **In scope:** what PF v2.0 is, what it tries to make easier to see, and how it is meant to be used (as orientation, not clearance).
- **Out of scope:** proofs of effectiveness; a certification regime; a complete ethical theory.
- **Notes:** conceptual; no internal cross-references by design.

Potentialism Framework (PF) is a proposed framework for making ethical judgment **more legible and revisable** in context—especially when people disagree, incentives distort reporting, or stakes make "good intentions" an unreliable proxy.

PF organizes attention around a small set of terms (defined later in the front matter): **potentials** (neutral capacities), **expressions** (what occurs in a moment), and **context** (the conditions shaping expression). It treats **will** as the trained capacity to regulate expression, and **ethics** as the cultivated skill of choosing expressions that are as mutually compatible as possible while reducing avoidable suffering and protecting dignity.

PF uses **compatibility** as an evaluative posture: assessing foreseeable effects on (a) avoidable suffering/harm, (b) dignity of awareness (as a ceiling principle), and (c) others' freedom to regulate their potentials within real constraints (to a reasonable degree). It treats **responsibility** as scaling with awareness and power/impact: higher awareness and higher impact carry more obligation to use will in line with ethical insight and to answer for effects.

PF is offered as a compass, not a rulebook: it provides concepts and practice scaffolds that can support inquiry, contestation, and revision, without claiming to replace domain evidence, governance, or technical assurance.

0.1.1 Bridge to the next section

The next section offers a reader map: where to start and how to navigate PF depending on your purpose.

CC BY–NC 4.0

0.2 Reader Map

- **In scope:** quick routes through PF for different reader goals.
- **Out of scope:** detailed argumentation (handled in the Parts).
- **Notes:** conceptual; navigation-heavy by design.

If you only read a few pages, these are common starting routes:

- **New to PF (fast orientation):** Abstract → Scope and limits → Mini glossary.
- **Quoting or reusing PF:** How to cite → License.
- **"Is PF validated / ready?":** Status and maturity → Validation roadmap.
- **Core concepts (the conceptual spine):** Core architecture.
- **Practical use without turning PF into compliance:** Operationalization + Misuse-resistance.
- **Metrics and measurement caution:** Measurement posture.
- **Cross-cultural / cross-domain translation:** Positioning and translation.

If you're reading PF in a high-stakes context, it can help to treat PF as supplementary orientation and keep the boundary posture visible (see Section 0.3) alongside the maturity posture (see Section 0.5).

0.2.1 Bridge to the next section

The next section sets PF's scope and limits: what PF is trying to do, what it does not claim, and where it hands off to other forms of evidence and governance.

CC BY–NC 4.0

0.3 Scope and limits

- **In scope:** what PF is for; what it is not for; how to avoid "PF-as-clearance" misreads; evidence posture and handoff posture.
- **Out of scope:** a governance regime; a certification standard; a universal maturity model; claims of empirical validation.
- **Notes:** conceptual; boundary-focused.

PF is a **conceptual proposal** with practice scaffolds. Its aim is to make ethical judgment in context easier to **inspect, contest, and revise**—especially when power, incentives, or institutional narratives would otherwise make reasoning opaque.

PF is **not**:

- a certification scheme ("PF-approved"),
- a substitute for domain evidence or technical assurance,
- a governance enforcement mechanism,
- a claim of empirical effectiveness.

PF can be useful as a shared vocabulary and a set of scaffolds for documenting how judgments were made. But PF does not replace the need for:

- domain methods and evidence (where empirical claims are in play),
- technical assurance and safety/security practices (where relevant),
- governance mechanisms with real authority (where accountability requires enforcement),
- translation work when PF crosses cultures, domains, or institutional settings (see Part VI).

0.3.1 Evidence posture

PF distinguishes between:

- **conceptual / normative claims** (the document's primary content), and
- **empirical / technical assurance claims**, which should be cited when made and treated as contestable inputs rather than moral clearance.

PF distinguishes between conceptual / normative claims and empirical / technical assurance claims. Where helpful, this distinction may be signaled with inline claim-basis cues (see Section 0.7.1).

CC BY–NC 4.0

When PF is used to support real decisions, a recurring failure mode is treating artifacts (templates, scores, narratives) as if they were proof. PF names and resists this drift in Part IV.

0.3.2 Bridge to the next section

The next section gives the preferred citation string and versioning practices, so PF references stay traceable across revisions.

CC BY–NC 4.0

0.4 How to cite PF v2.0

- **In scope:** recommended citation formats for PF v2.0; how to reference sections and versions so others can locate the exact wording you relied on.
- **Out of scope:** a mandatory citation style; claims that citation implies endorsement or approval.
- **Notes:** conceptual. Citations here are meant as **locators** for traceability across revisions.

When referencing PF v2.0, include:

- **Title:** *Potentialism / Potentials Framework (PF v2.0)*
- **Version identifier:** PF v2.0 (and build/date if available)
- **Section reference:** section number **and** section title

A minimal example:

> *Potentialism / Potentials Framework (PF v2.0)*, Section 2.5 "Fast/slow calibration," [DATE/BUILD TBD].

0.4.2 Citing adaptations (excerpt, translation, summary)

If you reuse or adapt PF, include:

- a note on **what changed** (e.g., "excerpted," "translated," "adapted," "summarized"),
- the **source version** you used,
- and where the adapted text can be found.

This helps separate PF's original wording from downstream interpretations.

0.4.3 Bridge to the next section

The next section states PF's status and maturity posture: what kind of document this is, and how to interpret "maturity" without turning it into a compliance ladder.

CC BY–NC 4.0

0.5 Status and maturity level

- **In scope:** maturity posture (proposed/evolving); how to interpret PF use as stakes rise; what "maturity" means in PF without implying tiers or certification.
- **Out of scope:** a universal maturity model; claims that PF is proven effective; policy directives about when PF "may" or "may not" be used.
- **Notes:** conceptual; status is disclosed to prevent "artifact-as-proof" misreads.

PF v2.0 is offered as a **proposed framework** designed for **traceable revision**. It does not present itself as empirically validated; instead, it frames validation and adoption as an open program of testing, critique, and revision (see Part VII).

0.5.1 What "maturity" means here

In PF's posture, "maturity" refers to whether a way of using the framework has been tested enough that its strengths, limits, and failure modes are becoming reasonably well understood—and whether that understanding is recorded in ways that can be contested and revised.

When PF use is generating real learning, it often shows up as:

- **traceability** (others can reconstruct what was done and why),
- **contestability** (disagreement has a pathway to matter),
- **revision visibility** (changes respond to critique and observed failure modes),
- **boundary discipline** (clarity about what PF cannot substitute for).

These are not a checklist or a ladder; they are signs that PF is being used as inquiry rather than theater.

0.5.2 Appropriate use by stakes and context

PF can be used in many contexts. As stakes rise—particularly when power/impact is high, harms could be severe, or dignity of awareness is plausibly at risk—users often find it wise to pair PF with:

- clearer documentation of what was known, assumed, and uncertain,
- explicit handoffs to domain evidence, technical assurance, and governance,
- a willingness to slow down, narrow scope, or defer when contestability is weak.

In such settings, "success" can include stopping early, reducing ambition, or declining to treat PF artifacts as decision authority (see Section 0.3; see Part IV).

CC BY–NC 4.0

0.5.3 What this status does and does not imply

In practice, PF is often used as:

- a conceptual map for attention and tradeoffs (see Part I),
- a set of practice scaffolds for legibility and revisability (see Part II),
- a measurement posture that treats instruments as learning supports, not verdicts (see Part III),
- a misuse-resistance stance that treats "artifact-as-proof" as a live risk (see Part IV),
- a validation roadmap that treats pilots as inquiry, not proof (see Part VII).

PF is typically less useful when treated as:

- a badge or certification regime,
- a universal scoring system that settles Compatibility,
- a universal cultural template that assumes its originating vocabulary travels unchanged (see Part VI).

0.5.4 Cultural and translation posture

PF is written from within particular linguistic and cultural conditions. Cross-cultural use is therefore treated as translation work: preserving Compatibility's criteria—avoidable harm/suffering, dignity of awareness as a ceiling principle, and others' freedom to regulate their potentials within real constraints—while expecting genuine disagreement about meanings and legitimacy (see Part VI).

0.5.5 Bridge to the next section

PF's maturity posture depends on definitions staying stable enough to be inspectable. The next section provides the mini glossary that anchors key terms used throughout PF.

CC BY–NC 4.0

0.6 Mini glossary

- **In scope:** reference meanings of key PF terms as used in this document; a stable anchor for internal consistency.
- **Out of scope:** broader philosophical debates about these words in general use; elaborations beyond what PF needs for consistent navigation.
- **Notes:** conceptual.

Consistent use of these terms helps keep PF's orientation stable across contexts and interpretations. When precision matters, it can help to use the meanings below rather than paraphrasing.

- **Potential** — a neutral underlying capacity of a system or being to generate patterns of experience or behavior when triggered. Sources of potentials may include biology/embodiment, learned habits, meanings/narratives/values, and extended systems/tools (including engineered components).
- **Expression** — a concrete, situated manifestation of one or more potentials in a specific moment and context.
- **Context** — the concrete configuration of conditions in which an expression takes place (e.g., relational field, power/impact, resources/constraints, incentives, norms/rules/instructions, awareness, and internal states).
- **Will** — the trained capacity to regulate how, when, and how intensely potentials are expressed.
- **Ethics** — the cultivated insight/skill of choosing expressions that are as mutually compatible as possible, while protecting dignity and reducing avoidable suffering. *(See "dignity of awareness" below.)*
- **Compatibility** — evaluating expressions relative to their foreseeable effects on: (a) avoidable suffering/harm, (b) dignity of awareness, and (c) others' freedom to regulate their potentials within real constraints (to a reasonable degree).
- **Awareness** — (operationally) the capacity to understand consequences, model self/others over time, and regulate expression. In PF, this refers to an underlying or realistically exercisable capacity, not merely to what is being successfully used in a given moment.
- **Dignity of awareness** — wherever awareness is present to a meaningful degree, dignity applies as a ceiling principle: no "systemic compatibility" can justify the destruction or humiliation of awareness beyond limits. Differences in awareness are **not** a ranking of basic worth; they primarily affect what is at stake, what protections are owed, and how responsibility scales.

CC BY–NC 4.0

- **Responsibility** — obligation proportional to awareness and power (impact) to use will in line with ethical insight, and to answer for effects.
- **Regulated responsibility (Awareness × Power)** — the higher the insight and the higher the impact, the more ethical weight is carried.

0.6.1 Bridge to the next section

The next section explains how PF v2.0 was written and what that process does and does not guarantee.

CC BY–NC 4.0

0.7 How this was written and limitations

- **In scope:** inputs and constraints; the role of model assistance; what this process can and cannot guarantee.
- **Out of scope:** a defense of conclusions; claims of empirical validation; a prescriptive workflow for others.
- **Notes:** conceptual; provenance is included to support traceability and interpretability.

PF v2.0 was produced as a revision of earlier PF/manifesto materials and feedback, with the goal of improving clarity, internal consistency, and misuse-resistance while keeping PF's compass posture intact (see Part IV). The process is described here so readers can better interpret the text's status and limits, and so future revisions can remain traceable rather than silent (see Section 0.4).

0.7.1 Inputs and constraints

This rewrite was guided by three core constraints:

- **Definition stability:** core terms are used with the mini glossary meanings (see Section 0.6).
- **Claim-basis cue:** where helpful, the text may use inline markers such as *[CONCEPTUAL]* to indicate a primarily conceptual / interpretive claim, and *[NEEDS CITATION]* to flag claims that lean empirical or assurance-like and require evidential support.
- **Compass posture:** the aim is to orient judgment under uncertainty, not to certify compliance or precompute outcomes (see Part II).

0.7.2 How the revision was done (descriptive)

In practice, the work moved through recurring phases: drafting for structure, review for rulebook tone and definition drift, readability polishing, and consistency checks against surrounding sections—sometimes paired with adversarial or "skeptical reader" passes to surface likely loopholes or misreads. This describes how this release was produced; it is not presented as a required method for reuse or future updates.

0.7.3 Human role and model assistance

PF v2.0 was produced with explicit human editorial ownership and model assistance.

CC BY–NC 4.0

- The human author/editor set aims and constraints, selected inclusions/exclusions, enforced definition stability, and made final decisions about revisions.
- Model assistance was used as a drafting and review tool under human direction (structure, clarity, tone checks, and misuse-oriented critique).

Model assistance changes the labor of drafting, not the status of the claims. Errors and blind spots remain possible—especially in edge cases, contested concepts, and culturally loaded language (see Part VI).

0.7.4 Known limitations

As a proposed framework, PF remains:

- **unvalidated in an empirical sense** (validation/adoption are treated as an open program; see Part VII),
- **selective by design** (not an exhaustive synthesis of traditions, standards, or domain risk methods; see Section 0.3),
- **culturally situated** (cross-context use requires translation work; see Part VI),
- **vulnerable to misuse under incentives** even when failure modes are named (see Part IV).

0.7.5 Bridge to the next section

The next section clarifies reuse rights and attribution expectations.

CC BY–NC 4.0

0.8 License

- **In scope:** reuse rights for this text; what attribution and change-marking typically include; how to avoid implying endorsement.
- **Out of scope:** legal advice; licensing for third-party materials quoted elsewhere; settling edge-case questions about what counts as "commercial."
- **Notes:** conceptual. This is a plain-language summary; the canonical license text governs. *[Creative Commons n.d.]*

This work is licensed under **CC BY–NC 4.0** (Creative Commons Attribution–NonCommercial 4.0 International). *[Creative Commons n.d.]*

This section summarizes the license for clarity; it is a legal condition of reuse, not an extension of PF's conceptual claims.

0.8.1 What you can do (license permissions)

The CC BY–NC 4.0 license permits reusers to:

- **Share** — copy and redistribute the material in any medium or format.
- **Adapt** — remix, transform, and build upon the material. *[Creative Commons n.d.]*

0.8.2 License terms (plain-language summary)

Reuse under CC BY–NC 4.0 is predicated on conditions including:

- **Attribution:** give appropriate credit, provide a link to the license, and indicate if changes were made—without suggesting the licensor endorses you or your use.
- **NonCommercial:** do not use the material for commercial purposes.
- **No additional restrictions:** do not apply legal terms or technological measures that restrict others from doing what the license permits. *[Creative Commons n.d.]*

For the canonical license text, see: https://creativecommons.org/licenses/by-nc/4.0/

0.8.3 Practical attribution and change-marking

Attribution can take many formats; for PF, a traceability-oriented format often helps (see Section 0.4). Common elements include:

- title of the work as it appears on the cited version,
- author/issuing organization (if applicable),
- version identifier and date (if applicable),
- where you accessed it,

CC BY–NC 4.0

- a short note if you made changes (e.g., “adapted,” “excerpted,” “reformatted,” “translated,” “summarized”).

If your reuse draws on only part of PF, it can help to cite the relevant section(s) so readers can find the exact wording you relied on (see Section 0.4).

0.8.4 Notices and warranty posture

- You do not need to comply with the license for elements that are in the public domain or where your use is permitted by an applicable exception or limitation. *[Creative Commons n.d.]*
- The license provides the material **as-is** and does not provide warranties; other rights (e.g., privacy, publicity, moral rights) may still constrain use. *[Creative Commons n.d.]*

0.8.5 Bridge to the next section

Licensing clarifies reuse rights; the next section provides an overview of major changes and motivations across versions.

CC BY–NC 4.0

0.9 Change log overview

- **In scope:** a high-level map of major changes from PF/Manifesto v1.1 to PF v2.0, and the motivations for those changes—primarily for readers migrating from v1.1.
- **Out of scope:** an exhaustive diff; claims that v2.0 is "better" in any validated sense; re-litigating underlying debates.
- **Notes:** conceptual; orientation-only. A detailed change log can be found in Appendix H.

PF v2.0 revises v1.1 materials with an emphasis on traceability, navigability, and misuse-resistance, while keeping core definitions stable (see Section 0.6) and keeping validation/adoption as an explicit open program (see Section 0.5; see Part VII).

0.9.1 Structural and routing changes

- **Expanded front matter**: added clearer boundary, citation/versioning, glossary, provenance, and licensing sections (see Sections 0.3–0.8).
- **Reorganized into Parts**: separated core architecture from practice scaffolds, measurement posture, misuse-resistance, conflict/repair, translation, and validation (see Part I–Part VII).
- **Navigation pins**: cross-references are used to help readers locate where topics are treated in depth, without turning prose into a procedure (see Section 0.4).

0.9.2 Posture shift (manifesto → compass)

v2.0 puts more weight on preventing two misreads:

- treating ethical language as certification ("PF-approved"), and
- treating procedures or artifacts as substitutes for judgment ("we did the steps, therefore it's fine").

This shows up as a generally softer cadence outside definition-locked lines, and as templates framed as optional scaffolds that scale with stakes and constraints (see Part II; see Part IV).

0.9.3 What is newly explicit (or more consolidated)

- **Definitions consolidated**: a mini glossary gathers definition-locked terms to reduce drift (see Section 0.6).

CC BY–NC 4.0

- **Compatibility not collapsed to a single metric**: Compatibility is treated as multi-lens and context-sensitive, especially where measurement and incentives can distort interpretation (see Part III; see Part IV).
- **Ceiling principle emphasis**: dignity of awareness is repeatedly treated as a ceiling constraint on what "systemic" reasoning can justify (see Part IV; see Part V).

0.9.4 Practice scaffolds and measurement posture

- **Operationalization added**: v2.0 develops lightweight artifacts (context snapshots, compatibility judgments, responsibility maps, decision records) as aids for legibility and contestability, not compliance (see Part II).
- **Measurement posture added**: v2.0 introduces a distinct stance toward instruments and metrics—learning supports rather than moral verdicts—with explicit attention to proxy failure and gaming (see Part III).

0.9.5 Misuse, translation, and validation as explicit programs

- **Misuse-resistance expanded**: v2.0 names failure modes such as compatibility-washing, substitution by declared intention/metrics/procedure, and capture/legitimation drift (see Part IV).
- **Translation made explicit**: cross-cultural and cross-domain use is treated as translation work by function, not one-to-one word transfer (see Part VI).
- **Validation and adoption roadmap added**: v2.0 frames pilots, criteria, documentation, and revision governance as an explicit program rather than an assumption of readiness (see Part VII; see Section 0.5).

0.9.6 Bridge to Part I

Front matter ends here. Part I begins the framework itself: the core architecture and how PF's key terms fit together in practice.

CC BY–NC 4.0

0.10 References (Front matter)

1 Creative Commons. n.d. “Attribution–NonCommercial 4.0 International (CC BY–NC 4.0).” Creative Commons. Accessed 2026-02-28. URL: https://creativecommons.org/licenses/by-nc/4.0/

CC BY–NC 4.0

1 Core architecture

Part I introduces the Potentialism Framework (PF) as a **compass**: a small set of concepts for orienting ethical judgment in real contexts without pretending to precompute answers. It defines the core vocabulary—**Potential**, **Expression**, **Context**, **Will**, **Ethics**, **Awareness**, **Compatibility**, **Responsibility**, and **Dignity of Awareness**—and shows how they relate.

The aim is to make ethical attention more legible: how inner capacities and outer constraints interact, how dignity sets ceilings on tradeoffs, and how responsibility scales with awareness and impact. This part is primarily conceptual: it offers mapping tools and cautionary guardrails rather than measurement claims or governance recipes. Part II then turns to practice: what PF outputs when used in concrete decisions and complex socio-technical systems.

CC BY–NC 4.0

1.1 Orientation: compass, not rulebook

- **In scope:** how to read PF as a compass (orientation, tradeoffs, thresholds), not a complete moral theory; what PF is trying to help you notice in real decisions.
- **Out of scope:** a compliance standard; a full meta-ethics; "PF settles every dilemma"; legal doctrine.
- **Notes:** conceptual. Key terms are used in their locked mini-glossary sense.

PF is not a catalogue of moral rules. It is an **orientation tool** for evaluating **expressions**—what actually happens in a concrete moment—of underlying **potentials** (capacities) within a specific **context** (the surrounding conditions that shape meaning and effects). It uses three lenses of **Compatibility**: **harm** (avoidable suffering), **dignity** (the dignity of awareness ceiling), and **agency** (others' freedom to regulate their potentials within genuine constraints).

A central guardrail is a ceiling principle: **appeals to "systemic benefits" are treated as insufficient, by themselves, to justify severe destruction or humiliation of awareness beyond limits** (see Section 1.9). PF is intended to support ethical reflection—especially in complex situations shaped by power, information asymmetries, or vulnerability—without replacing judgment.

PF in one minute (quick orientation) (see Appendix F)

PF helps you evaluate what is being expressed, in what context, who is affected, and how to move toward compatibility—while treating dignity as a hard constraint on certain tradeoffs. It supports judgment under uncertainty; it does not replace it.

PF treats ethics primarily as a **skill of navigation**, not a list of final answers. In general, it aims to improve the *questions* we ask before we act—especially when situations are complex, contested, or shaped by power.

1.1.1 What PF is trying to change

Many ethical conversations collapse into identity judgments ("good people" vs. "bad people") or trait labels ("this drive is evil," "that capacity is virtuous"). PF shifts focus from *what something is* to *how it is expressed* in a specific situation. This can enable critique without freezing beings into a single story, and it preserves the possibility that a capacity can be expressed more compatibly over time.

CC BY–NC 4.0

PF also orients ethical attention by **awareness** rather than species membership: ethical consideration tracks capacities for experience, vulnerability, and harm, and **responsibility scales with awareness and impact**. This is not meant to erase human obligations; it is meant to reduce unexamined "human-only" assumptions in cases where awareness and impact are the ethically relevant features.

1.1.2 How PF orients judgment

PF evaluates expressions in context rather than passing verdicts on essences. It orients reflection around five recurrent questions:

Five guiding questions (scan-friendly)

1. Which potentials seem active here (what capacities are in play)?
2. What expression is actually happening (not only what is intended)?
3. What is the context—relationships, power/impact, constraints, incentives, norms/instructions, relevant internal states, and (where helpful) intentions?
4. Who is affected, and what are their awareness, vulnerability, and capacity to respond?
5. How compatible is this expression, in this context, with the three lenses—harm, dignity, and agency—and are constraints genuinely fixed or partly shaped by earlier decisions and power arrangements?

These questions do not guarantee a single universal answer. They are meant to make reasoning more explicit, more reviewable, and less vulnerable to rationalization.

As a parallel check—especially when stakes are high—PF also asks how responsibility scales with the acting agent(s)' **awareness and power/impact**.

Practical tools that implement this posture appear in Part II (see Sections 2.2 and 2.4). They are designed to support action under uncertainty without turning PF into a rulebook.

1.1.3 Compass posture: guidance without imprisonment

PF is written as a compass because beings with awareness and will are often not fully predictable. Attempts to imprison ethical judgment in fixed rule lists can become ethically risky—especially under power. A framework that tries to precompute every outcome can become a cage: it can be misused as moral cover while neglecting context, impact, and power dynamics.

CC BY–NC 4.0

PF offers orienting commitments rather than rigid axioms, encouraging judgment that is:

- **Context-aware** (sensitive to constraints and power, not only intentions)
- **Provisional yet accountable** (open to correction as consequences become clearer, while still acting on the best available understanding and leaving a trace for review)
- **Responsibility-centered** (justification burden tends to grow with awareness and impact)
- **Ceiling-bounded** (the dignity ceiling constrains what tradeoffs can be treated as ethically legible)

Common misreadings (short clarifiers)

- *Not relativism: context-sensitive, but bounded by harm, dignity, and agency.*
- *Not compliance: it does not replace law or safety standards; it helps ask whether "mere compliance" is ethically adequate (see Section 0.3).*
- *Not intention-only: intentions matter, but sit alongside power, incentives, and foreseeable effects.*
- *Not human-only: it orients attention by awareness and vulnerability where relevant.*
- *Not "anything goes": judgments remain accountable to effects, power, and responsibility scaling.*

1.1.4 Where this posture continues

For PF's formal scope, limits, and deference posture, see Section 0.3. For how to apply PF in practice, begin with Part II (Sections 2.2 and 2.4).

1.1.5 Bridge to the next section

With the compass posture set, Part I begins with the most basic unit PF tracks: **Potentials**—neutral capacities that can later be expressed in many ways.

CC BY–NC 4.0

1.2 Potentials

- **In scope:** Potential as a neutral underlying capacity of a being/system; how to notice potentials (including in tools and institutions) without moralizing them by default.
- **Out of scope:** metaphysical debates about "potentiality" in general; detailed psych/neuro claims; declaring which potentials are "good" or "bad" independent of context.
- **Notes:** conceptual; examples are illustrative, not empirical claims.

A **potential** is a neutral underlying capacity of a system or being. PF uses "neutral" to mean **not yet a moral verdict**—not "harmless" and not "excused." PF postpones moral verdicts about a capacity until we consider **how it is expressed**, by whom, and under what conditions.

A simple example: **physical strength** is a potential. It can be expressed as protection, care, skilled work, intimidation, or violence—depending on context, regulation, and impact. PF's question is not "is strength good?" but "how is it being expressed here, and with what effects?"

PF evaluates at the level of **expression in context**. Still, latent potentials matter in one practical sense: if a high-impact actor or system holds a capacity that could be expressed destructively, it can affect how we think about safeguards, oversight, and responsibility—even before it is activated.

Instead of asking "Is this trait good or bad?" PF asks: **How is this capacity being expressed here, and how does it land on the three compatibility lenses—harm, dignity, and agency—given the constraints that actually apply?** PF also invites attention to "constraints": some are close to material or safety limits; others are shaped by incentives, defaults, and power.

PF's aim is not to deny stable dispositions. It is to keep moral attention anchored to what is *actionable*: what is being expressed and what can be regulated or redesigned.

1.2.1 Potentials are neutral: divergent expression pathways

The examples below are meant to make neutrality concrete. They are **not** virtues, and they are not stable rankings. The same capacity can branch into more compatible or less compatible expressions depending on timing, power, alternatives, and context.

- **Assertiveness / agency**
  Often more compatible: stating boundaries; advocating for someone with less

CC BY–NC 4.0

power.
Often less compatible: domination; silencing; escalating conflict when de-escalation is available.

- **Sensitivity to threat**
  Often more compatible: caution in high-risk environments; noticing early warning signs.
  Often less compatible: chronic suspicion; scapegoating; pre-emptive aggression.
- **Curiosity**
  Often more compatible: learning; asking better questions; exploring under uncertainty.
  Often less compatible: intrusion; violating privacy; "experimentation" without consent.
- **Empathy / social attunement**
  Often more compatible: care; perspective-taking; repairing relationships.
  Often less compatible: emotional overreach; manipulation; enabling harmful behavior.
- **Conscientiousness / order**
  Often more compatible: reliability; follow-through; safety-oriented diligence.
  Often less compatible: rigidity; punitive rule-enforcement; suppressing necessary change.
- **Need for recognition / status**
  Often more compatible: healthy ambition; taking responsibility publicly.
  Often less compatible: performative virtue; coercive status games; sacrificing others for prestige.

A practical implication is that "strength vs. weakness" language can drift into moral labeling, when the more useful distinction is often **fit**: which expressions are compatible *here*, and what regulation (Will) or context design would prevent predictable failure modes *[Grant and Schwartz 2011; Kaiser and Kaplan 2009; Niemiec 2019]*.

1.2.2 A layered view: where potentials come from

PF treats potentials as multi-sourced and layered. This is a **conceptual lens**, not a claim that human behavior cleanly separates into modules; in real life, layers interact.

1. **Biology / embodiment (temperament, affect, energy)**
   Baseline sensitivities, arousal patterns, and embodied constraints *[Rothbart 2011; McCrae and Costa 1997]*.

CC BY–NC 4.0

2. **Learned patterns and habits**
   Repeated responses that become automatic, especially under stress or time pressure *[Wood and Neal 2007]*.
3. **Meanings, narratives, and values**
   Identity stories and meaning-making frames that shape what feels permissible *[McAdams and McLean 2013]*.
4. **Extended systems and tools (including engineered components)**
   Institutions, interfaces, incentives, norms, and technologies that expand or steer what a person—or an organization—can do *[Clark and Chalmers 1998]*.
   Note: tools and systems can have potentials **independent of any particular user** (e.g., a platform's amplification potential, a surveillance system's targeting potential). A user can trigger or channel them, but the capacity can also be designed, constrained, or redirected at the system level.

This layered framing supports PF's broader aim: evaluate expressions in context while resisting fixed moral labels.

1.2.3 "Soft skills" and identity-upgrade pressure (brief note)

"Soft skills" often mixes **trainable behaviors** with **personality expectations**, which can slide from skill-building into identity demands *[Duckworth and Yeager 2015; Deming 2017]*. PF's ethical focus stays on improving expressions (and contexts that shape them), rather than requiring status-laden "become a different person" upgrades as a condition of dignity.

Training pathways and institutional learning belong primarily in Part II's progression materials (see Section 2.7), where PF can address practice without turning Part I into a program manual *[Baldwin and Ford 1988; Durlak et al. 2011; Heckman and Kautz 2012; Kautz et al. 2014; Roberts, Walton, and Viechtbauer 2006; Roberts et al. 2017]*.

1.2.4 Potentials vs. types and traits

PF is compatible with describing dispositions as traits (e.g., Big Five research), as long as traits are not treated as moral essences and are interpreted through expression-in-context *[John, Naumann, and Soto 2008; Tett and Burnett 2003]*. PF is also compatible with using "types" as informal heuristics, as long as they are not reified into fixed identities or used to excuse harmful expressions.

Popular typology language can become socially sticky (e.g., "this is just who I am"), which is one reason PF keeps returning to **potentials + expressions + context** rather than identity labels *[16Personalities 2025]*.

CC BY–NC 4.0

1.2.5 Bridge to the next section

Once we can name a potential, the next step is to notice its **expression**: what actually manifests in a particular moment.

CC BY–NC 4.0

1.3 Expressions

- **In scope:** what PF means by *expression*; why PF evaluates expressions rather than essences; what commonly counts as an evaluable expression across humans, social systems, and AI-enabled systems.
- **Out of scope:** a complete taxonomy of expression-types; measurement/psychometrics; legal or compliance guidance.
- **Notes:** primarily conceptual here; later sections add lightweight tools and review practices.

An **expression** is a concrete, situated manifestation of one or more potentials in a specific moment and context. "Situated" is literal: expressions happen within relationships, power/impact, constraints, incentives, norms, and internal states that shape meaning and effects.

1.3.1 Why PF evaluates expressions (not essences)

PF concentrates evaluation where it can be most **concrete and reviewable**: at the level of what happened in context—rather than verdicts about a person's essence or a capacity's moral "nature."

This aligns with a well-known pattern in psychology and organizational research: behavior varies with situations, and observers often over-attribute outcomes to stable character while under-weighting constraints and incentives *[Ross 1977; Mischel and Shoda 1995; Tett and Burnett 2003]*.

In PF terms: potentials are not moral verdicts; expressions are where compatibility questions become actionable.

1.3.2 What counts as an expression

Because PF applies across humans, social systems, and AI-enabled systems, "expression" is intentionally broad—but still concrete. In practice, an expression can include:

- **Actions and omissions:** what is done, and what is foreseeably not done *when a specific action was realistically available and reasonably expected in that role* (as distinct from "not doing everything").
- **Speech and signaling:** what is said, implied, promised, threatened, or normalized—especially when it shapes others' options.

CC BY–NC 4.0

- **Design and deployment choices:** defaults, interfaces, incentive structures, access controls, escalation paths, monitoring choices that tend to steer behavior in practice *[Thaler and Sunstein 2008]*.
- **Delegations and enforcement:** who is authorized, who is excluded, what is punished or rewarded, and how rules are applied in practice.

The key is not the category; it is whether we can point to a specific event-pattern in time: "this policy was deployed with these defaults," "this decision was made under these constraints," "this model output was used in this workflow."

A practical heuristic (reviewability without surveillance)
If you can describe an event-pattern so a second party could review it—even imperfectly—PF generally treats it as an evaluable expression. "Review" can be personal reflection, a trusted peer, or a confidential process; it is not an argument for invasive surveillance that violates legitimate privacy [Nissenbaum 2010].

1.3.3 Expression is not "intention only" (and not "outcome only")

PF does not reduce ethics to outcomes alone. Intentions, beliefs, and motives often matter—especially as signals of awareness, care, and future reliability. PF treats them as part of context, which can change what was reasonably foreseeable and what alternatives were realistically available.

At the same time, good intention is not a blanket permission slip. A well-meant expression can still be judged less compatible if it foreseeably increases avoidable harm, violates the dignity ceiling, or unnecessarily constrains others' agency.

1.3.4 Expressions are often composite

In real situations, a single "action" often bundles multiple potentials (care + anxiety + status-seeking) expressed through multiple channels (speech, timing, resource allocation) and shaped by systems (norms, platforms, incentives). PF therefore evaluates many expressions as **composite**: which effects dominate, which harmful elements were avoidable given awareness and alternatives, and what re-channeling (via Will and/or context design) would make future expressions more compatible.

CC BY–NC 4.0

1.3.5 Responsibility shows up at the expression level

Responsibility scales with awareness and power/impact. Expressions are where this becomes concrete: as impact grows and as access to relevant information increases, the expectation generally rises to regulate expression more carefully and to offer reasons for tradeoffs.

PF does not assume equal agency; context can sharply narrow options. This is why PF encourages attention to whether constraints are genuinely material/safety-related or partly maintained by avoidable design choices and power dynamics.

1.3.6 Bridge to the next section

Expressions are never "in a vacuum." The next section unpacks **Context** as the configuration that shapes what an expression becomes and what it costs.

CC BY–NC 4.0

1.4 Context

- **In scope:** what PF means by *context*; recurring dimensions of context; why context changes how expressions are interpreted and evaluated.
- **Out of scope:** a full theory of social causality; legal standards for "reasonable" behavior; comprehensive political analysis of power.
- **Notes:** primarily conceptual here.

Context is the concrete configuration of conditions in which an expression takes place. PF emphasizes context because the same outward act can carry different meaning and different effects across conditions. A widely used framing is that behavior reflects both person and environment *[Lewin 1936]*, and observers commonly over-attribute outcomes to "who someone is" while under-weighting situational constraints *[Ross 1977; Mischel and Shoda 1995]*.

PF's stance is not "context explains everything." It is: **context is part of what makes expressions ethically legible.**

1.4.1 Why context matters in PF

Context helps PF avoid collapsing complex situations into simple labels; it supports "what happens next?" reasoning (since effects often depend on incentives, power gradients, and constraints—not only intention); and it helps keep responsibility proportional to what was realistically knowable and doable in a role.

Context is therefore a tool for explanation and fair evaluation: a higher-resolution description of what is actually being expressed, by whom, under which conditions, and with what feasible alternatives—without treating context as automatic exoneration.

1.4.2 The recurring dimensions of context

This list is not a completeness test. It is a set of recurring lenses that often change compatibility judgments.

1 **Relational field**
Who is involved; roles and histories; trust, dependency, conflict; vulnerability and care dynamics.

2 **Power and impact**
Asymmetries in authority, resources, information, credibility, and exit options; the scale and reach of the expression.

CC BY–NC 4.0

3 **Resources and constraints**
Time, money, tools, access, safety conditions, and the option-set (what alternatives were actually available). PF treats constraints as ethically relevant—and also asks whether some are partly shaped by avoidable design or policy choices.

4 **Incentives and pressures**
Rewards, penalties, metrics, deadlines, social approval, threats, and hidden costs. Incentives can steer expressions in recurring directions *[Thaler and Sunstein 2008]*.

5 **Norms, rules, and instructions**
Explicit rules (laws, policies, procedures), implicit norms (culture, taboo, expectations), and instructions from authorities or systems.

6 **Available information and situational awareness**
What information was accessible or reasonably expected for someone in that role; what warnings or disclosures existed; what was salient. This is not a demand for omniscience.

7 **Internal states (moment-level)**
Fatigue, stress, fear, grief; attention and cognitive load.
Clarifier: "biology/embodiment" (as a source of potentials) is the deeper baseline; "internal states" are the nearer-term conditions that can fluctuate hour-to-hour and narrow regulation capacity in the moment.

8 **Temporal trajectory / history**
The lead-up: prior interactions, accumulated harms, precedents, and "why the option-set looks like this now." History often changes what an expression means, what repair is possible, and what future risks are likely.

These dimensions interact. A design choice (defaults) can amplify incentives, which can shape internal states, which can reduce regulation capacity (Will), shifting expressions toward lower compatibility.

1.4.3 "Real constraints" and the definition-capture risk

PF keeps "real constraints" language because constraints are real and ethically relevant. At the same time, PF treats constraint-claims as a **site of scrutiny**, not a free pass.

Constraints are rarely binary. Some are close to physical or safety limits; others emerge from social arrangements, resource allocation, institutional history, and power. Many are mixtures (e.g., "budget" often reflects both scarcity and priorities). PF encourages asking:

- *Real for whom?* (whose safety, whose costs, whose options)
- *Chosen by whom?* (policy choices, defaults, incentive design)
- *Maintained by what?* (institutions, dependencies, lock-in, enforcement)

CC BY–NC 4.0

This is not a claim that constraints are usually fake. It is a prompt to separate material/safety limits from constraints created or maintained by avoidable choices, and to notice when "reasonableness" may be defined mainly by those with greater power.

1.4.4 Context in AI-enabled and socio-technical systems

In AI-enabled settings, context often includes the surrounding workflow and institutional environment: who deploys the system, for what purpose, with what defaults, what oversight, and what fallback options. Two identical model outputs can function as different expressions depending on whether they are

- a suggestion in a human review loop,
- an automated decision with no appeal path,
- a tool used by a high-power institution against a low-power individual.

This is one reason PF treats design and deployment choices as expressions: socio-technical context shapes downstream effects and the distribution of constraints *[Trist and Bamforth 1951]*.

1.4.5 Bridge to the next section

If context shapes expression from the outside, **Will** names a capacity that can shape expression from within—how regulation and training can change what gets expressed over time.

CC BY–NC 4.0

1.5 Will

- **In scope:** what PF means by *Will*; Will as a capacity; why Will matters for shaping expressions over time; how Will interacts with context and responsibility.
- **Out of scope:** clinical advice; detailed neuroscience; a complete training program; legal "capacity" doctrines.
- **Notes:** primarily conceptual here.

Will is the trained capacity to regulate **how, when, and how intensely** potentials are expressed. In PF, Will is not a moral verdict and not a guarantee of good outcomes; it is a regulatory capacity that can support more careful expressions in context.

Because Will is not directly observable, PF often **infers** it from patterns of regulation over time—not from outcomes alone. A regulated expression may look "mild" for many reasons (support, low stakes, luck). PF's concern is whether a system shows **adjustment**—modulating intensity, timing, or channel of expression as context changes.

A recurring confusion is to treat Will as sheer force ("push harder") or as having the "right" desires. PF uses Will in a narrower sense: **the capacity to modulate expression**, especially under pressure.

1.5.1 Will as capacity and as skill

PF treats Will as skill-like: it can strengthen (or degrade), and it can be supported by practice and feedback over time. Regulatory capacity also appears graded: many systems show more or less flexible regulation depending on development, learning history, stress load, and context.

In practice, signs of regulation capacity can include:

- **Flexible modulation:** varying intensity, timing, or channel rather than executing a fixed response.
- **Inhibition / delay:** sometimes pausing or suppressing a dominant impulse when conditions change.
- **Updating from feedback:** adjusting future expression based on consequences or signals.

In humans, these functions are often studied under labels like executive functions / cognitive control *[Miyake et al. 2000; Diamond 2013]*. Comparative work suggests some nonhuman animals show forms of inhibitory control and flexible regulation, with degree and reliability varying by species and context *[MacLean et al. 2014]*.

CC BY–NC 4.0

PF is cautious about attributing Will where indicators of flexible regulation are unclear. For plants and very simple organisms, debates about cognition and agency remain active; PF does not settle them here and uses Will-language cautiously, tied to observable regulation patterns and feedback sensitivity *[Taiz et al. 2019; Calvo et al. 2020]*.

Compass implication: when Will-capacity is minimal or uncertain, PF tends to emphasize **context design** and external supports/constraints over exhortations to "use more will."

1.5.2 Will interacts with context

Will does not operate in a vacuum. Context shapes what is feasible, salient, rewarded, risky, and safe. Sustained threat, deprivation, overload, or coercion can narrow regulation capacity and reduce what is realistically doable in the moment *[Shonkoff et al. 2012]*. In such cases, improving outcomes may depend more on changing context than on demanding greater individual control.

1.5.3 Will, awareness, and responsibility

Responsibility scales with awareness and power/impact. Will is one pathway through which that scaling becomes concrete: as awareness and impact increase, the expectation that an actor can regulate expression tends to increase—while still recognizing that coercion, deprivation, and institutional constraints can sharply limit what is feasible.

Will and awareness interact without being identical: greater awareness expands what is foreseeable and what alternatives are thinkable, while Will concerns the capacity to regulate expression in light of what is foreseeable and feasible.

1.5.4 Training Will (when capacity exists)

PF treats Will as trainable because regulation can improve through practice, supports, and feedback in many humans *[Diamond 2013]*. This is not a universal promise; evidence for "training" effects in adults is mixed and improvements are often context-specific, with limited transfer to unrelated domains *[Diamond and Ling 2016; Melby-Lervåg, Redick, and Hulme 2016]*. PF therefore treats training as one lever among others, often alongside context design.

Illustrative supports that can increase reliable regulation include:

- stability supports (reducing avoidable volatility),
- pause-and-check habits in higher-stakes contexts,
- re-channeling (redirecting a potential into a more compatible expression),

CC BY–NC 4.0

- perspective expansion (making consequences more foreseeable, consistent with role and access).

1.5.5 Will in AI-enabled and socio-technical systems

PF uses "Will" most naturally for agents with internal regulation. In AI-enabled systems, the closest counterparts may be **engineered regulation mechanisms**: constraints, guardrails, monitoring, feedback loops, rate limits, escalation policies, and human-in-the-loop governance that modulate how, when, and how intensely a system acts over time.

Because regulation in socio-technical systems is often distributed, PF's compass move is to map the regulatory chain: what is regulated by humans, by institutions, by software constraints, and by incentives and defaults. This helps avoid both over-attributing agency to machines and under-attributing responsibility to deployers.

A useful prompt remains: **what is doing the regulation—where, and under whose control?**

1.5.6 Common failure modes and misuses

- **Will as blame:** using "lack of will" to ignore coercion, deprivation, disability, or power asymmetries.
- **Will as domination:** equating "strong will" with imposing outcomes on others rather than regulating one's own expression.
- **Will as purity test:** treating willpower as moral ranking rather than a capacity with context-dependent limits.
- **Will as substitute for ethics:** improving self-control without changing orientation can make harmful expression more efficient.

1.5.7 Bridge to the next section

Because Will can be trained, PF frames **Ethics** less as a single choice and more as a cultivated skill of choosing compatible expressions over time.

CC BY–NC 4.0

1.6 Ethics

- **In scope:** what PF means by *Ethics*; what it means to call ethics an insight/skill; how ethics relates to Will, context, and compatibility.
- **Out of scope:** complete moral theory; legal compliance standards; technical safety specifications for AI systems.
- **Notes:** primarily conceptual here.

Ethics is the cultivated insight/skill of choosing expressions that are as mutually compatible as possible, while protecting dignity and reducing avoidable suffering. In PF, ethics is not a static code that mechanically outputs answers; it is a cultivated way of seeing and choosing that may improve with reflection, feedback, and institutional learning.

PF distinguishes **criteria** from **capacity**: **Compatibility** names the evaluation lens; **Ethics** names the cultivated ability to apply that lens to real expressions in context—under uncertainty, power differences, and tradeoffs.

(Terminology note) In everyday speech, "morality" often refers to rules and prohibitions, while "ethics" can refer to reflective judgment. PF does not depend on that distinction; it uses **Ethics** in the specific, glossary-locked sense above.

1.6.1 What "insight" means in PF

Calling ethics an insight does not mean "whatever feels right." In PF, insight points to practical, accountable understanding—an improved ability to notice what matters and to choose accordingly:

- noticing interdependence (power, incentives, constraints),
- tracking the dignity ceiling,
- anticipating second-order effects across time,
- generating alternatives (not treating conflicts as one-move games),
- staying revisable (updating as consequences reveal blind spots),
- giving reasons (making tradeoffs challengeable rather than hidden behind slogans).

This overlaps with traditions emphasizing practical wisdom—applying ethical understanding to particulars, not only reciting rules *[Aristotle 2002, bk. VI; Schwartz and Sharpe 2010]*. PF borrows this practical idea without importing a full virtue theory.

CC BY–NC 4.0

1.6.2 Compatibility-seeking and long-horizon stability (sometimes)

PF does not assume compatibility-seeking "wins" in every case. It claims something narrower: in some interdependent settings, compatibility-seeking can also be instrumentally wise over time, because it can preserve trust, reduce costly cycles of conflict, and expand feasible joint action.

Several research traditions offer models consistent with this **under certain conditions**:

- **Repeated interaction and reciprocity:** In repeated games, strategies supporting conditional cooperation can outperform purely exploitative strategies by sustaining mutual benefit and deterring defection [Axelrod 1984; Trivers 1971; Nowak 2006].
- **Collective action and durable institutions:** Communities can maintain cooperation around shared resources when rules and incentives support monitoring, fairness, and conflict resolution [Ostrom 1990].
- **Mutual-gains negotiation:** In bargaining, focusing on underlying interests and expanding the space of options can sometimes produce agreements that leave both sides better off than positional "win/lose" bargaining [Fisher, Ury & Patton 1991].

These findings apply most cleanly when actors expect repeated interaction, have some capacity to enforce commitments, and are not trapped in purely one-shot predation dynamics. PF's ethical orientation does not depend on compatibility paying off; it remains anchored by harm reduction and the dignity ceiling even when compatibility is costly.

1.6.3 Ethics, Will, and "power without orientation"

Will and ethics play different roles:

- **Will** concerns regulation (modulating expression).
- **Ethics** concerns orientation (choosing relative to compatibility, dignity, and avoidable harm).

This matters because stronger regulation can amplify outcomes in either direction. Technical competence or discipline can make harmful expression more effective if orientation is incompatible. PF treats ethics-as-insight as a check on capability without compatibility: not anti-competence, but competence guided by the right evaluative lens.

1.6.4 Ethics as a cultivated skill (without a rulebook)

PF calls ethics a skill because it involves capacities that can be practiced and, in favorable conditions, strengthened—individually and collectively. Skill-like components often include

CC BY–NC 4.0

perception (seeing what is happening in context), imagination (generating alternatives), judgment under uncertainty, and coordination (shaping norms and processes so compatible expressions become easier).

This aligns with long-standing work on moral development and ethical judgment as involving capacities supported by education, practice, and institutions *[Kohlberg 1984; Rest 1986]*. PF does not endorse a single developmental theory; it uses the general idea that ethical reliability is not only rule knowledge.

1.6.5 Ethics in AI-enabled and socio-technical systems

In socio-technical settings, ethical "insight" often appears as design and governance choices: what is optimized, what is constrained, who can appeal decisions, and how harms are detected and corrected. PF does not replace technical safety work; it orients how goals and constraints are selected, justified, audited, and revised for compatibility *[NIST 2023; OECD 2019; IEEE 2021; ISO/IEC 2023]*.

A useful compass question is not only "whose goals are being served?" but also: **how are benefits and burdens distributed across affected parties over time?**

1.6.6 Common misunderstandings to avoid

- **Ethics as compliance:** "ethical" = "whatever rules currently say," even when rules misalign with dignity or harm reduction.
- **Ethics as purity:** ethics as identity label rather than practice.
- **Ethics as conflict-avoidance:** compatibility is not passivity; it can include firm boundaries and resistance.
- **Ethics as mere prudence:** compatibility as strategy for personal success rather than evaluative stance anchored by dignity and harm reduction.

1.6.7 Bridge to the next section

To keep "skill" from becoming vague, **Compatibility** provides the evaluation lens PF uses when asking: "compatible with what, and for whom?"

CC BY–NC 4.0

1.7 Compatibility

- **In scope:** what PF means by *Compatibility*; the three components; why Compatibility is evaluated across time; how "real constraints (to a reasonable degree)" functions as a scrutiny point rather than a loophole.
- **Out of scope:** a universal scoring system; legal standards for "reasonable" behavior; a complete theory of social power; a technical risk model for AI systems.
- **Notes:** primarily conceptual. Practical checklists/protocols are in Part II.

Compatibility is PF's central evaluation lens: evaluating expressions relative to their foreseeable effects on (a) avoidable suffering/harm, (b) dignity of awareness, and (c) others' freedom to regulate their potentials within real constraints (to a reasonable degree).

"Reasonable degree" functions as both:

- an **epistemic posture** (use the information reasonably available for the role, given time/resources; state uncertainty), and
- a **practical posture** (the depth of inquiry and burden of justification tend to scale with stakes, impact, and reversibility).

PF treats Compatibility as a **multi-lens** evaluation, not a single score. Strength on one lens does not automatically cancel failure on another.

1.7.1 Compatibility is evaluated across time

Compatibility is not only about how an expression lands now. An expression can look locally compatible while setting conditions that make future incompatibilities more likely—through retaliation, incentive shifts, precedent, lock-in, or the slow narrowing of agency.

Because foresight is limited, PF is cautious with reasoning that treats present violations of dignity or agency as acceptable means justified mainly by speculative future benefits—especially under uncertainty *[Knight 1921]*. PF therefore emphasizes attention to plausible downstream effects paired with explicit uncertainty and revisability.

1.7.2 The three components of Compatibility

PF keeps the components distinct because collapsing them tends to create blind spots.

1.7.2.1 Avoidable suffering / harm

Compatibility asks whether an expression foreseeably increases avoidable suffering or harm—relative to realistic alternatives in that context. "Avoidable" is not perfectionism; it

CC BY–NC 4.0

points to preventable harm given the option-set and constraints that were reasonably present.

PF treats harm as broader than direct physical injury. It can include psychological and social harms (e.g., intimidation, coercive control, humiliation, systematic exclusion) and harms that propagate through relationships or institutions over time *[Feinberg 1984; Parfit 1984]*.

At the same time, PF is cautious about collapsing harm into "anything I dislike." Disagreement or offense can be emotionally real but is not automatically treated as harm that justifies coercion *[Mill 1859; Feinberg 1985]*.

(Practical note, kept conceptual) In practice, harm judgments often depend on **severity**, **scope**, and **reversibility** across time *[Parfit 1984]*. Part II provides compact prompts to support consistent triage.

1.7.2.2 Dignity of awareness (ceiling principle)

PF treats dignity of awareness as a ceiling: some "solutions" become ethically illegible once they require severe degradation, humiliation, or destruction of awareness beyond limits.

PF does not define the ceiling by a numerical score. It relies on **operational anchors**—paradigm cases widely treated as clear violations—to calibrate what "beyond limits" can look like in practice (e.g., torture and cruel, inhuman, or degrading treatment) *[United Nations 1984]*. Approaching these anchors in mechanism or intent often signals the need to slow down, broaden review where possible, and prefer reversible options.

This ceiling is one reason PF is wary of collapsing ethics into aggregate optimization: "systemic success" is treated as insufficient, by itself, to justify crossing the dignity ceiling *[Rawls 1971; Scanlon 1998]*.

1.7.2.3 Others' freedom to regulate their potentials (within real constraints)

Compatibility includes whether an expression foreseeably restricts others' freedom to regulate their potentials within real constraints. In practice, this includes asking whether a claimed constraint is truly binding for affected parties, and whether a restriction goes beyond what those constraints plausibly justify.

Restrictions on freedom are treated as high-salience moves. Narrowing others' agency can sometimes be justified to prevent more severe harms, but PF is cautious when restrictions are justified mainly by lower-salience impacts such as offense or ideological discomfort. Where restriction is proposed, PF encourages considering proportionality and less-

CC BY–NC 4.0

restrictive alternatives—an intuition echoed in necessity/proportionality/least-restrictive-means traditions *[United Nations Economic and Social Council 1984; Childress et al. 2002]*.

1.7.3 Compatible, incompatible, and mixed outcomes

PF often uses "compatible" and "incompatible" as shorthand, but expects many real cases to be mixed:

- **Compatible:** reduces avoidable harm, respects the dignity ceiling, preserves agency within constraints.
- **Incompatible:** foreseeably increases avoidable harm, violates the dignity ceiling, or needlessly narrows agency.
- **Mixed:** reduces one harm while increasing another; protects dignity while narrowing agency; helps some groups while harming others.

In mixed cases, Compatibility is meant to provide structure for deliberation (naming costs, stating uncertainty, tracking who bears which burdens), not to guarantee a clean answer.

1.7.4 Where the "how-to" lives

Part I keeps Compatibility conceptual. The compact checklist and step-by-step decision flow live in Part II (Sections 2.2 and 2.3), along with documentation practices that keep reasoning transparent and revisable (Section 2.4).

1.7.5 Common misuses to watch for

- **Compatibility as "be nice":** compatibility is not passivity; it can include firm boundaries and resistance.
- **Compatibility as optimization:** treating it as a single score invites sacrificing dignity or agency for aggregate gains.
- **Compatibility as excuse:** "constraints" can be used to rationalize preventable harm; PF treats constraint-claims as ethically relevant and contestable.
- **Compatibility as certainty:** overconfident forecasting can turn speculative future benefits into permission for present harm *[Knight 1921]*.

1.7.6 Bridge to the next section

Compatibility depends on what can be understood and anticipated; the next section clarifies **Awareness** as the capacity to model consequences and regulate expression over time.

CC BY–NC 4.0

1.8 Awareness

- **In scope:** PF's operational meaning of *Awareness*; why Awareness matters for foreseeability, Will, and Responsibility; how PF distinguishes Awareness from "consciousness"; how PF treats non-human and artificial forms of Awareness under uncertainty.
- **Out of scope:** a full theory of consciousness; a definitive test for sentience; legal standards for personhood/rights; a technical benchmark for "AI awareness".
- **Notes:** primarily conceptual.

Awareness (operationally, in PF) is the capacity to understand consequences, model self/others over time, and regulate expression. In practice, this capacity depends partly on access to relevant information: opacity, deprivation, and information asymmetry can materially limit awareness even when a system or agent has underlying modeling capacity. PF uses an operational definition because Compatibility evaluates foreseeable effects, and responsibility scales with an agent's ability to foresee and self-regulate over time.

In PF, Awareness is about a being's capacity, not its social status, formal education, or what it is successfully expressing in a given moment. A being may have relevant awareness-capacity even when fear, deprivation, manipulation, fatigue, or structural constraint limit its present use. For that reason, Awareness should not be read as a basis for ranking beings by worth; in PF it matters mainly for foreseeability, forms of protection, and Responsibility scaling.

This is a scope choice, not a metaphysical claim.

1.8.1 Awareness and consciousness (why PF separates the questions)

In philosophy of mind, "consciousness" is often used to mean felt experience—what it is like to be in a state—rather than (or in addition to) the ability to perceive, model, and control behavior *[Block 1995; Chalmers 1996]*. PF's term Awareness does not settle that debate.

PF's use is operational: if a system demonstrates a sufficient capacity to track consequences across time, model itself and others, and regulate expression, then its behavior becomes ethically legible in a way that matters for PF's purposes—whether those capacities arise from biological experience, engineered mechanisms, or sophisticated simulation.

This separation matters because different ethical questions can hinge on different kinds of evidence:

CC BY–NC 4.0

- **Responsibility, training, and dignity:** operational Awareness is directly relevant.
- **Suffering-like harms:** some harms may depend on felt experience. Our access to felt experience (in humans, animals, or machines) is limited and contested, which becomes ethically relevant under uncertainty *[Birch et al. 2021; Butlin et al. 2023]*.

1.8.2 Why Awareness matters in PF

Awareness changes what Compatibility can reasonably ask of an expression, and it changes how responsibility is assigned:

1. **Foreseeability expands** with greater Awareness.
2. **Regulation becomes possible**: where Awareness is higher, Will can function as trained regulation.
3. **Responsibility scales with Awareness × Power**: where impact is high and Awareness is low, PF tends to shift responsibility toward designers, operators, and institutions shaping deployment.
4. **Dignity scales with Awareness and constrains tradeoffs**: severe degradation or destruction of Awareness carries exceptional ethical weight (especially under uncertainty).

1.8.3 Non-human and artificial forms of Awareness

PF does not assume Awareness is human-only. It treats Awareness as a capacity that can vary by degree and form, including non-human animals and, in principle, engineered systems as their modeling and self-regulation capabilities advance.

PF also treats uncertainty as ethically relevant:

- If a system shows non-trivial operational Awareness (planning across time, self-monitoring, uncertainty handling), interventions that would severely degrade or destroy those capacities become higher-stakes.
- When it is unclear whether a system has felt experience (or an analogue), PF tends to treat severe, irreversible harm as requiring unusually strong justification and to prefer reversible, reviewable options where feasible *[Birch et al. 2021; Butlin et al. 2023]*.

PF aims to avoid both anthropomorphic inflation and deflationary dismissal.

CC BY–NC 4.0

1.8.4 Practical prompts (not a test)

These are thinking aids, not checkboxes:

- consequence modeling across time,
- self-modeling (limits, uncertainty, failure modes),
- other-modeling (others' incentives/constraints),
- regulation capacity (inhibition, redirection),
- uncertainty signaling,
- impact awareness (adjusting behavior when stakes rise).

1.8.5 Bridge to the next section

As Awareness grows, so does what is at stake; **dignity of awareness** adds a ceiling principle that constrains what tradeoffs can justify.

CC BY–NC 4.0

1.9 Dignity of awareness

- **In scope:** PF's meaning of dignity of awareness; why PF treats it as a ceiling principle; how it relates to Awareness and Compatibility; how PF reasons under uncertainty; how to proceed when a situation seems to force a dignity-limiting tradeoff.
- **Out of scope:** full moral theory of dignity; complete rights/personhood framework; legal standards for detention/war/punishment; technical tests for consciousness/sentience.
- **Notes:** conceptual.

PF treats **dignity of awareness** as a ceiling principle: **appeals to "systemic compatibility" are insufficient, by themselves, to justify severe destruction or humiliation of awareness beyond limits**. The ceiling resists a recurring failure mode: treating awareness-bearing beings as exchangeable units in a calculation, where sufficiently large benefits elsewhere are used to justify domination, humiliation, or erasure *[Kant 1785; Rawls 1971; Scanlon 1998]*.

This ceiling does not eliminate tradeoffs in the world. It changes what counts as an ethically legible tradeoff: when a proposal depends on degrading or destroying Awareness, PF treats the moral burden as exceptionally high.

1.9.1 Dignity scales with Awareness, and still functions as a ceiling

PF holds that dignity scales with awareness: for beings with richer and more temporally extended Awareness, more can be at stake when that Awareness is coerced, humiliated, or destroyed. "Scales with" signals gradation; "ceiling" signals that some forms of destruction or humiliation become ethically illegible beyond limits once Awareness is present to a meaningful degree.

This should not be read as a hierarchy of basic worth within aware beings. In PF, "scales with" means that the **form, depth, and stakes** of dignity-relevant harm may vary with the depth and kind of awareness, not that more educated, more articulate, or higher-status beings are more valuable. A person's social rank, profession, intelligence, or present performance is not a measure of superior dignity.

PF does not specify a numeric threshold. Instead, it treats ceiling risk as rising with:

- plausibility and degree of Awareness,
- severity and irreversibility of the intervention,

CC BY–NC 4.0

- whether the intervention targets humiliation, domination, or destruction (rather than temporary constraint for protection).

1.9.2 Awareness, consciousness, and why PF separates the questions

PF distinguishes operational Awareness from debates about consciousness as felt experience *[Block 1995; Chalmers 1996]*. This matters because:

- dignity can matter even when suffering is uncertain, and
- outsiders often lack decisive access to what a system "really feels," if anything—especially for novel artificial systems.

Under uncertainty, PF treats severe degradation as higher-stakes precisely because the downside may be hard to undo *[Birch et al. 2021; Butlin et al. 2023]*.

1.9.3 When "systemic benefits" conflict with dignity

PF anticipates situations where decision-makers claim dignity-limiting action is necessary to prevent larger harms. PF does not automatically dismiss such claims, but treats them as high-burden moves that call for explicit reasoning:

- name the ceiling risk plainly (avoid euphemisms),
- search for options that avoid targeting awareness for degradation,
- if a ceiling-adjacent action still appears necessary, acknowledge moral cost and narrow scope,
- bias toward reversibility, repair, and learning,
- broaden review when time allows.

These themes resonate with necessity/proportionality/least restrictive means traditions *[United Nations Economic and Social Council 1984; Childress et al. 2002]*.

1.9.4 Dignity under uncertainty for non-human and artificial systems

PF balances two errors:

- premature dismissal ("it's just simulation") as justification for severe harm,
- uncritical attribution from superficial resemblance.

As severity and irreversibility rise—and confidence about Awareness or felt experience falls—PF typically treats wider scrutiny as appropriate, without demanding paralysis.

CC BY–NC 4.0

1.9.5 Bridge to the next section

With dignity as a constraint, PF turns to **Responsibility**: how obligation scales with what an agent can understand and what they can affect.

CC BY–NC 4.0

1.10 Responsibility

- **In scope:** PF's meaning of Responsibility; why it scales with Awareness and Power/impact; how it behaves in chained decisions and institutions; what refusal and escalation mean in PF's compass posture; how to think about responsibility when information is missing or uncertainty is high.
- **Out of scope:** full legal liability theory; complete professional codes; full institutional design recipes.
- **Notes:** conceptual.

Responsibility is obligation proportional to awareness and power (impact) to use will in line with ethical insight, and to answer for effects. PF emphasizes this because Compatibility is evaluated in context and across time, and Awareness makes action ethically legible.

PF orients responsibility judgments by looking at:

1. **Capacity:** what the actor could understand and regulate (Awareness; Will).
2. **Reach:** what the actor could affect (Power/impact in context).
3. **Feasible levers:** what the actor could realistically do differently from their position, given time, information access, and constraints.

Responsibility is not primarily a label ("good" or "bad person"). It is a relationship between capacities, reach, and feasible levers—plus willingness to be answerable in proportion to those realities.

1.10.1 Responsibility is scalable, not binary

PF treats responsibility as graded rather than on/off:

- minimal Awareness is usually required for responsibility to attach meaningfully to regulation,
- as Awareness increases, justification and learning become more meaningful,
- as Power/impact increases (fast, wide, durable, hard-to-reverse), the burden of care and review tends to increase.

PF does not require precise measurement. In many settings, coarse bands (low/medium/high) are enough to orient mapping.

CC BY–NC 4.0

1.10.2 Delegation does not automatically erase responsibility

PF distinguishes delegating tasks from delegating ethical responsibility. In chains of action, delegation can distribute work—but it can also create "many-hands" dynamics where everyone touches an outcome and no one feels answerable *[Thompson 1980]*.

Responsibility is not removed by delegation alone. An actor or organization can outsource execution while still carrying responsibility for what they authorize, what safeguards they accept or reject, and how they respond when warning signs appear *[Arendt 1963]*.

1.10.3 Refusal and non-compliance as responsibility moves (within constraints)

Responsibility can include the capacity to **not do** certain things—where an actor has real discretion.

When an agent can recognize that an action is plausibly incompatible—especially because it risks crossing dignity limits or causing severe avoidable harm—PF treats refusal, delay, or constraint-seeking as potentially meaningful expressions of Will.

This is not framed as heroism or a universal demand. PF recognizes that coercive contexts can make refusal costly or unavailable. In such cases, PF shifts weight toward those with greater Power to create safe escalation channels and protect dissent, and toward best-feasible alternatives for constrained actors (raising concerns, narrowing scope, documenting risks, seeking protected advice).

1.10.4 Escalation when information is missing (orientation prompts)

PF does not assume perfect knowledge. But some situations call for seeking missing information, routing decisions through higher-awareness processes, adding safeguards, or considering a pause—especially when consequences may be wide, durable, or irreversible.

Self-orientation prompts for higher-impact roles:

- Who can actually see the relevant risks?
- Who has authority to slow down, stop, or add safeguards?
- What pressures and incentives are shaping silence or speed?
- Who bears the downside if uncertainty resolves badly?

These are compass prompts, not a compliance checklist.

CC BY–NC 4.0

1.10.5 Responsibility includes repair and learning, not only blame

Responsibility includes answering for effects in a forward-looking way: naming ethical costs without euphemism, narrowing further harm, repairing where possible, and learning how to avoid returning to the same corner—bounded by feasibility and proportionality.

1.10.6 Bridge to the next section

Because power and awareness are often distributed across systems, PF introduces **regulated responsibility** as a mapping habit for complex socio-technical chains.

CC BY–NC 4.0

1.11 Regulated responsibility: awareness x power

- **In scope:** PF's compact mapping tool—regulated responsibility (Awareness × Power); how to use it without pretending it is a metric; how it helps prevent "responsibility evaporation" in complex systems.
- **Out of scope:** full measurement theory for Awareness or Power; legal liability doctrine; detailed governance mechanisms.
- **Notes:** conceptual; citations point to established discussions about responsibility under complexity and bounded rationality.

PF uses a simple compass tool: **regulated responsibility (Awareness × Power)**. This is not a calculator. It is a mnemonic for keeping ethical weight attached to places where understanding and impact concentrate—especially when decisions are distributed across people, institutions, and tools.

A nearby theme in technology ethics is that modern action can scale in reach and irreversibility, expanding what it is reasonable to treat as responsibility-bearing *[Jonas 1984]*. PF expresses this as a mapping habit usable across domains.

1.11.1 Why "Awareness × Power" (and why the "×")

PF treats responsibility as scaling with both:

- **Awareness:** capacity to understand consequences over time and regulate expression,
- **Power/impact:** the reach, speed, durability, and reversibility of what an actor can set in motion.

The "×" is not literal arithmetic. It marks an interaction: high values on either axis change the ethical posture, and high values on both create a distinct situation. It also avoids a common misread: **low Awareness does not "zero out" responsibility**. Low-Awareness / high-impact configurations are higher-risk and often shift substantial responsibility toward those who design, authorize, supervise, or route impact through low-awareness parts of the system.

PF does not assume we can precisely measure either axis. Human judgment is bounded and organizations operate under uncertainty *[Simon 1957]*. The formula is directional: "Where are the peaks of Awareness? Of impact? Are they aligned, or misaligned in ways that raise risk?"

CC BY–NC 4.0

1.11.2 Not a metric: use bands, thresholds, and explicit uncertainty

To avoid fake precision, PF relies on:

- coarse bands (low/medium/high),
- treating both axes as uncertain estimates (uncertainty can justify slowing down in higher-impact settings),
- using thresholds: in low-stakes contexts, mapping adds little; in high-stakes contexts, even moderate uncertainty can warrant escalation.

1.11.3 Avoidable ignorance is not neutral

Responsibility assessments do not depend only on what an actor happens to know at a moment. In many contexts, they also consider whether the actor had a reasonably accessible opportunity—given stakes and real costs of learning—to reduce uncertainty before acting.

The compass point is not "everyone becomes an expert." It is that, **when stakes are high and an epistemic step is reasonably accessible**, responsibility often includes at least a minimal information-seeking move: basic inquiry, consulting expertise, or escalation. Honest "I don't know" can be real; it is not always a stable ethical resting place when not-knowing predictably externalizes harm.

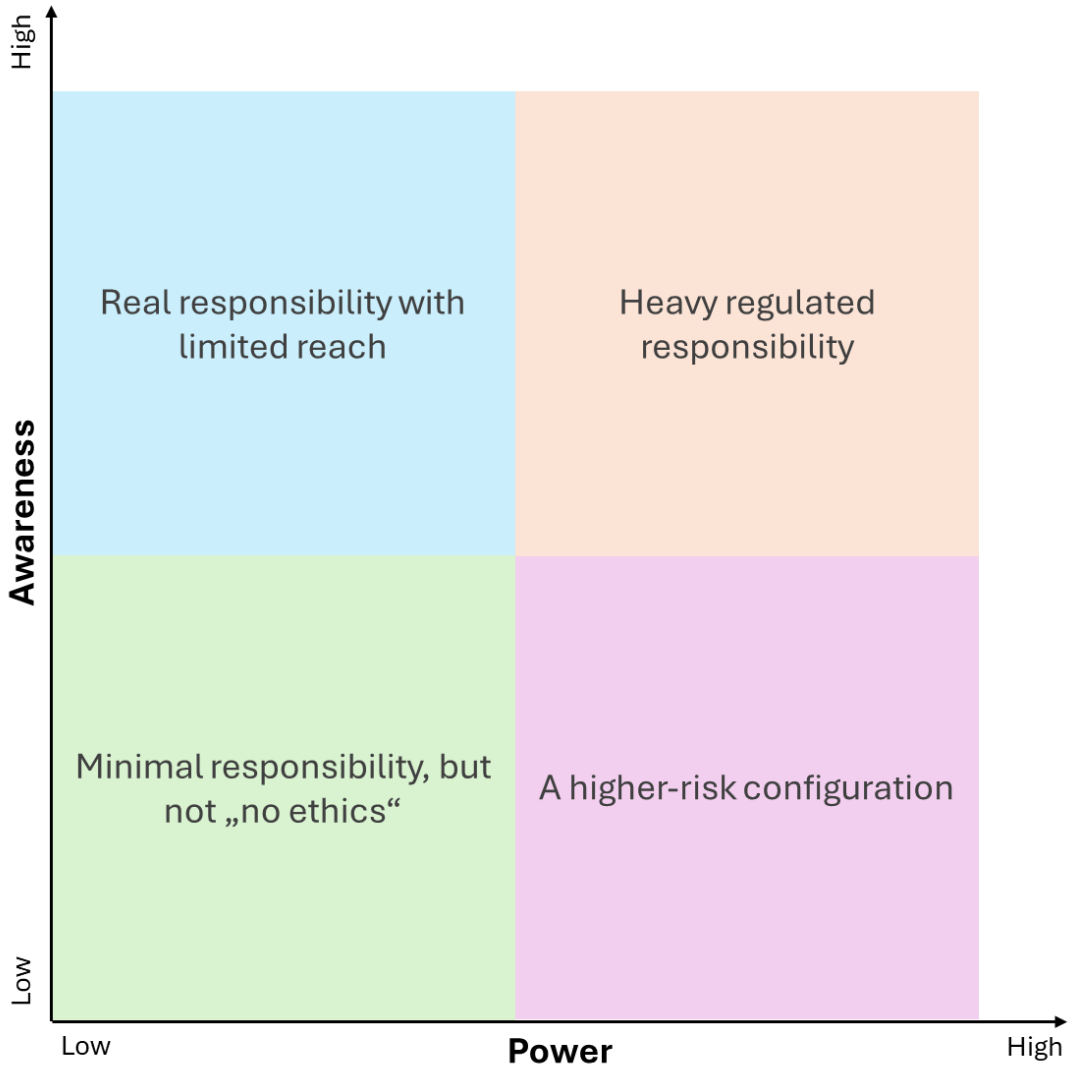
Some accounts treat ignorance as potentially culpable when it reflects avoidable failures of inquiry *[Rosen 2003]*. Work on willful ignorance analyzes cases where agents deliberately forgo inquiry to preserve deniability *[Zimmerman 2020]*. Experimental and social-psychological work suggests people may sometimes prefer ignorance when it is easy to maintain, and group settings can diffuse responsibility unless it is made explicit *[Dana, Weber, and Kuang 2007; Grossman and van der Weele 2017; Darley and Latané 1968; Latané, Williams, and Harkins 1979]*.

PF treats "reasonable expectation" as bounded: limited bandwidth, trauma, burnout, or lack of access can substantially reduce what is fair to demand. The aim is to avoid rewarding avoidable ignorance as a moral exit—especially where impact is high and burdens of not-knowing fall on others.

1.11.4 A quadrant sketch (heuristic illustration only)

This quadrant is not a moral taxonomy of persons. It is an illustration of how responsibility considerations often shift across four notional scenarios—while still checking for hidden constraints and indirect influence.

CC BY–NC 4.0



1.11.4.1 High Awareness, high Power: heavy regulated responsibility

Often calls for more care, justification, and restraint—frequently expressed as “slow mode”: gather information, consult, test assumptions, and be willing to pause when severe harm or dignity risks are plausible *[Jonas 1984; Weick and Sutcliffe 2015]*.

1.11.4.2 High Awareness, low Power: real responsibility with limited reach

Ethical weight is real, but ability to affect outcomes is constrained. Typical moves include warning, guidance, refusal, and escalation.

1.11.4.3 Low Awareness, high Power: a higher-risk configuration

Large impact driven by actors/components that cannot reliably understand consequences is higher-risk. PF often treats much responsibility as shifting toward those who grant,

CC BY–NC 4.0

maintain, or route impact through low-awareness parts of the system—designers, deployers, supervisors, institutions. In AI contexts, this connects to the "responsibility gap" discussion *[Matthias 2004]*.

1.11.4.4 Low Awareness, low Power: minimal responsibility, but not "no ethics"

Responsibility is limited, but context still matters; responsibility can shift toward those who design or control the environment.

1.11.5 Socio-technical mapping: preventing responsibility from evaporating

In real systems, Awareness and impact are distributed across individuals, institutions, tools, and interfaces. PF treats responsibility mapping as a practical application: identify where decision influence and relevant understanding sit, and—where feasible—make answerability explicit.

Two well-studied failure modes motivate this:

- diffusion through bureaucracy ("many hands") *[Thompson 1980]*,
- agency laundering ("the system decided") *[Rubel, Castro, and Pham 2019]*.

For structural harms from many ordinary actions, PF is directionally consistent with accounts that distinguish blame from forward-looking responsibility to change unjust structures *[Young 2011]*.

1.11.6 Guardrails against two common misreadings

- **"Awareness becomes authority."** Awareness increases responsibility to regulate ethically; it does not grant entitlement to dominate.
- **"Responsibility is just outcome-based blame."** Outcomes matter, but PF keeps process responsibility in view, acknowledging moral luck *[Nagel 1979; Williams 1981]* and emphasizing reasons-responsiveness and ownership of decision mechanisms *[Fischer and Ravizza 1998]*.

1.11.8 Bridge to the next section

Before Part II becomes practical, the next section briefly situates PF alongside major ethical traditions to help readers orient the vocabulary they've just learned.

CC BY–NC 4.0

1.12 A short map to existing traditions

- **In scope:** a navigation aid—how PF's core commitments resemble, diverge from, and extend several ethical traditions, without claiming equivalence.
- **Out of scope:** full history of ethics; comprehensive interpretation debates; "PF proves tradition X wrong."
- **Notes:** conceptual; citations are canonical anchors, not claims of consensus.

This section is intentionally brief. PF is not a replacement for major traditions; it is a compass that can be carried into them. The goal is to situate PF's terms—Ethics, Responsibility, Compatibility, Will, Awareness, and dignity of awareness—relative to familiar maps.

PF is also not defined as human-specific. Its core terms are stated at the level of beings and systems that express potentials in context. Where Awareness is present, PF's dignity ceilings and responsibility mapping become relevant.

1.12.1 Consequentialist family (utilitarian and related)

PF overlaps with consequentialist attention to effects through Compatibility's harm lens *[Mill 1863; Parfit 2011]*. Where it differs from strongly aggregative optimization is its dignity ceiling: severe humiliation or destruction of awareness is treated as ethically illegible beyond limits *[Rawls 1971; Scanlon 1998]*.

PF therefore treats consequentialist reasoning as important but incomplete: attention to suffering and benefits matters, but is regulated by dignity limits and by responsibility proportional to awareness and impact.

1.12.2 Deontological family (rights, duties, respect)

PF has a family resemblance to deontological concerns about strong constraints and respect, especially where dignity ceilings are in view *[Kant 1785]*. PF does not present ethics primarily as rule-following alone; it is explicitly context-dependent and skill-based. Duties and rights can be important tools for making boundaries legible and enforceable, without being the whole of ethical judgment.

A nearby bridge is contractarian work emphasizing what we can justify to one another under pluralism *[Rawls 1971; Scanlon 1998]*.

CC BY–NC 4.0

1.12.3 Virtue ethics and cultivation traditions

PF aligns with cultivation traditions in its emphasis on training: Will as regulation capacity and Ethics as cultivated insight/skill *[Aristotle 2002]*. PF adds explicit analytic handles—Potential → Expression → Context and Compatibility—so cultivation stays connected to structural constraints and downstream effects.

1.12.4 Care ethics and relational traditions

PF's emphasis on relational context, dependency, vulnerability, and burden distribution aligns with care-ethical concerns *[Gilligan 1982; Tronto 1993]*. PF links these to regulated responsibility: when impact concentrates, PF treats this as calling for stronger care and stronger justification—while keeping "reasonable expectation" bounded by real constraints so care is less likely to become a moral weapon.

1.12.5 Existentialism and "ethics of freedom" traditions

PF resonates with themes that agency is situated and that "role boundaries" are not always sufficient justification for foreseeable harm—especially under high impact *[Sartre 1946; Beauvoir 1947]*. PF adds explicit dignity ceilings and responsibility mapping across socio-technical chains.

1.12.6 Capability and freedom-oriented traditions

PF's agency lens (others' freedom to regulate potentials within constraints) resembles capability approaches focusing on what people are effectively able to do and be *[Sen 2009; Nussbaum 2011]*. PF's twist is keeping "freedom" connected to capacities (Potentials; Will) and power realism (responsibility scaling), not only formal rights.

1.12.7 Beyond human-centered ethics (condensed)

PF is compatible with approaches that extend moral attention beyond adult humans:

- **Animal ethics:** where sentience and suffering are central, PF's harm lens and dignity ceilings remain live constraints *[Singer 1975; Regan 1983]*.
- **Ecological/biocentric ethics:** PF can help keep "what is harmed, and at what scale" connected to context, power, and irreversibility, while being explicit about where Awareness-based ceilings do and do not apply *[Taylor 1986; Naess 1973]*.
- **Information and machine ethics:** PF keeps two questions distinct: (i) does a system plausibly have Awareness that triggers dignity ceilings? and (ii) regardless, who holds Awareness and impact in the surrounding human system such that

CC BY–NC 4.0

accountability does not disappear into "the system"? *[Moor 2006; Floridi 2013; Gunkel 2018]*

1.12.8 Responsibility-for-power traditions (technology ethics, governance)

PF's regulated responsibility mapping aligns with the view that modern technologies amplify reach and irreversibility, expanding the domain where foresight and accountability matter *[Jonas 1984]*. It also fits work on responsibility under distributed action ("many hands") and sociotechnical agency laundering *[Thompson 1980; Rubel, Castro, and Pham 2019]*, while remaining a compass posture rather than legal doctrine.

1.12.9 What PF adds (without replacing the traditions)

PF aims to add:

1. **A shared language for "inner" and "outer" ethics:** Potential, Expression, Context, Will.
2. **A multi-constraint evaluation lens:** Compatibility (harm, dignity ceiling, agency).
3. **A responsibility mapping habit:** Regulated responsibility (Awareness × Power) to keep accountability attached to real influence and understanding.

1.12.10 Bridge to Part II

Part I has established PF's core terms and compass posture. Part II starts from a practical question: **what does PF output when used in a real decision?**

CC BY–NC 4.0

1.13 References (Part I)

1. 16Personalities. 2025. “Our Framework.” *16Personalities*. Accessed December 23, 2025. https://www.16personalities.com/articles/our-theory
2. Arendt, Hannah. 1963. *Eichmann in Jerusalem: A Report on the Banality of Evil*. New York: Viking Press.
3. Aristotle. 2002. *Nicomachean Ethics*. Translated by Sarah Broadie and Christopher Rowe. Oxford: Oxford University Press.
4. Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books.
5. Baldwin, Timothy T., and J. Kevin Ford. 1988. “Transfer of Training: A Review and Directions for Future Research.” *Personnel Psychology* 41 (1): 63–105.
6. Beauvoir, Simone de. 1947. *The Ethics of Ambiguity*. New York: Philosophical Library.
7. Birch, Jonathan, et al. 2021. *Review of the Evidence of Sentience in Cephalopod Molluscs and Decapod Crustaceans*. London: UK Department for Environment, Food & Rural Affairs.
8. Block, Ned. 1995. “On a Confusion about a Function of Consciousness.” *Behavioral and Brain Sciences* 18 (2): 227–247.
9. Butlin, Patrick, et al. 2023. “Consciousness in Artificial Intelligence: Insights from the Science of Consciousness.” *arXiv* preprint arXiv:2308.08708.
10. Calvo, Paco, et al. 2020. “Plants Are Intelligent, Here’s How.” *Annals of Botany* 125 (1): 11–28. https://doi.org/10.1093/aob/mcz155.
11. Chalmers, David J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.
12. Childress, James F., Ruth R. Faden, Ruth D. Gaare, Nancy E. Kass, Anna C. Larson, R. Alta Charo, and others. 2002. “Public Health Ethics: Mapping the Terrain.” *Journal of Law, Medicine & Ethics* 30 (2): 170–178. https://doi.org/10.1111/j.1748-720X.2002.tb00384.x.
13. Clark, Andy, and David Chalmers. 1998. “The Extended Mind.” *Analysis* 58 (1): 7–19. https://doi.org/10.1093/analys/58.1.7.
14. Dana, Jason, Roberto A. Weber, and Jason Xi Kuang. 2007. “Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness.” *Economic Theory* 33: 67–80. https://doi.org/10.1007/s00199-006-0153-z.
15. Darley, John M., and Bibb Latané. 1968. “Bystander Intervention in Emergencies: Diffusion of Responsibility.” *Journal of Personality and Social Psychology* 8 (4): 377–383. https://doi.org/10.1037/h0025589.

CC BY–NC 4.0

16. Deming, David J. 2017. “The Growing Importance of Social Skills in the Labor Market.” *Quarterly Journal of Economics* 132 (4): 1593–1640. https://doi.org/10.1093/qje/qjx022.
17. Diamond, Adele. 2013. “Executive Functions.” *Annual Review of Psychology* 64: 135–168. https://doi.org/10.1146/annurev-psych-113011-143750.
18. Diamond, Adele, and Daphne S. Ling. 2016. “ Conclusions about interventions, programs, and approaches for improving executive functions that appear justified and those that, despite much hype, do not,” *Developmental Cognitive Neuroscience* 2016 Apr;18:34-48. DOI: 10.1016/j.dcn.2015.11.005
19. Duckworth, Angela L., and David S. Yeager. 2015. “Measurement Matters: Assessing Personal Qualities Other Than Cognitive Ability for Educational Purposes.” *Educational Researcher* 44 (4): 237–251. https://doi.org/10.3102/0013189X15584327.
20. Durlak, Joseph A., Roger P. Weissberg, Allison B. Dymnicki, Rebecca D. Taylor, and Kriston B. Schellinger. 2011. “The Impact of Enhancing Students’ Social and Emotional Learning: A Meta-Analysis of School-Based Universal Interventions.” *Child Development* 82 (1): 405–432. https://doi.org/10.1111/j.1467-8624.2010.01564.x.
21. Feinberg, Joel. 1984. *Harm to Others*. Vol. 1 of *The Moral Limits of the Criminal Law*. New York: Oxford University Press.
22. Feinberg, Joel. 1985. *Offense to Others*. Vol. 2 of *The Moral Limits of the Criminal Law*. New York: Oxford University Press.
23. Fischer, John Martin, and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.
24. Fisher, Roger, William Ury, and Bruce Patton. 1991. *Getting to Yes: Negotiating Agreement Without Giving In*. New York: Penguin Books.
25. Floridi, Luciano. 2013. *The Ethics of Information*. Oxford: Oxford University Press.
26. Gilligan, Carol. 1982. *In a Different Voice: Psychological Theory and Women’s Development*. Cambridge, MA: Harvard University Press.
27. Grant, Adam M., and Barry Schwartz. 2011. “Too Much of a Good Thing: The Challenge and Opportunity of the Inverted U.” *Perspectives on Psychological Science* 6 (1): 61–76. https://doi.org/10.1177/1745691610393523.
28. Grossman, Zachary, and Joël J. van der Weele. 2017. “Self-Image and Willful Ignorance in Social Decisions.” *Journal of the European Economic Association* 15 (1): 173–217. https://doi.org/10.1093/jeea/jvw001.
29. Gunkel, David J. 2018. *Robot Rights*. Cambridge, MA: The MIT Press.

CC BY–NC 4.0

30. Heckman, James J., and Tim Kautz. 2012. “Hard Evidence on Soft Skills.” *Labour Economics* 19 (4): 451–464. https://doi.org/10.1016/j.labeco.2012.05.014.
31. IEEE. 2021. *IEEE Standard Model Process for Addressing Ethical Concerns during System Design*. IEEE Std 7000-2021. Piscataway, NJ: IEEE.
32. ISO/IEC. 2023. *Artificial Intelligence—Guidance on Risk Management*. ISO/IEC 23894:2023. Geneva: International Organization for Standardization / International Electrotechnical Commission.
33. John, Oliver P., Laura P. Naumann, and Christopher J. Soto. 2008. “Paradigm Shift to the Integrative Big Five Trait Taxonomy: History, Measurement, and Conceptual Issues.” In *Handbook of Personality: Theory and Research*, edited by Oliver P. John, Richard W. Robins, and Lawrence A. Pervin, 3rd ed., 114–158. New York: Guilford Press.
34. Jonas, Hans. 1984. *The Imperative of Responsibility: In Search of an Ethics for the Technological Age*. Chicago: University of Chicago Press.
35. Kant, Immanuel. 1785. *Groundwork of the Metaphysics of Morals*. Translator: Mary Gregor, Editors: Karl Ameriks and Desmond M Clarke. Cambridge University Press
36. Kaplan, Robert E., and Robert B. Kaiser. 2009. “Stop Overdoing Your Strengths.” *Harvard Business Review*.
37. Kautz, Tim, James J. Heckman, Ron Diris, Bas ter Weel, and Lex Borghans. 2014. *Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success*. DOI: 10.3386/w20749
38. Knight, Frank H. 1921. *Risk, Uncertainty and Profit*. Boston: Houghton Mifflin.
39. Kohlberg, Lawrence. 1984. *Essays on Moral Development, Volume II: The Psychology of Moral Development*. San Francisco: Harper & Row.
40. Latané, Bibb, Kipling Williams, and Stephen Harkins. 1979. “Many Hands Make Light the Work: The Causes and Consequences of Social Loafing.” *Journal of Personality and Social Psychology* 37 (6): 822–832. https://doi.org/10.1037/0022-3514.37.6.822.
41. Lewin, Kurt. 1936. *Principles of Topological Psychology*. New York: McGraw-Hill.
42. MacLean, Evan L., Brian Hare, Charles L. Nunn, et al. 2014. “The Evolution of Self-Control.” *Proceedings of the National Academy of Sciences* 111 (20): E2140–E2148. https://doi.org/10.1073/pnas.1323533111.
43. Matthias, Andreas. 2004. “The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata.” *Ethics and Information Technology* 6 (3): 175–183. https://doi.org/10.1007/s10676-004-3422-1.
44. McAdams, Dan P., and Kate C. McLean. 2013. “Narrative Identity.” *Current Directions in Psychological Science* 22 (3): 233–238. https://doi.org/10.1177/0963721413475622.

CC BY–NC 4.0

45. McCrae, Robert R., and Paul T. Costa Jr. 1997. “Personality Trait Structure as a Human Universal.” *American Psychologist* 52 (5): 509–516. https://doi.org/10.1037/0003-066X.52.5.509.
46. Melby-Lervåg M, Redick TS, Hulme C. *Working Memory Training Does Not Improve Performance on Measures of Intelligence or Other Measures of "Far Transfer"*: Evidence From a Meta-Analytic Review. Perspect Psychol Sci. 2016 Jul;11(4):512-34. doi: 10.1177/1745691616635612.
47. Mill, John Stuart. 1859. *On Liberty*. London: John W. Parker and Son.
48. Mill, John Stuart. 1863. *Utilitarianism*. London: Parker, Son, and Bourn.
49. Mischel, Walter, and Yuichi Shoda. 1995. “A Cognitive-Affective System Theory of Personality: Reconceptualizing Situations, Dispositions, Dynamics, and Invariance in Personality Structure.” *Psychological Review* 102 (2): 246–268. https://doi.org/10.1037/0033-295X.102.2.246.
50. Miyake, Akira, Naomi P. Friedman, Michael J. Emerson, Andrew H. Witzki, and Amy Howerter. 2000. “The Unity and Diversity of Executive Functions and Their Contributions to Complex ‘Frontal Lobe’ Tasks: A Latent Variable Analysis.” *Cognitive Psychology* 41 (1): 49–100. https://doi.org/10.1006/cogp.1999.0734.
51. Moor, James H. 2006. “The Nature, Importance, and Difficulty of Machine Ethics.” *IEEE Intelligent Systems* 21 (4): 18–21.
52. Nagel, Thomas. 1979. “Moral Luck.” In *Mortal Questions*, 24–38. Cambridge: Cambridge University Press.
53. Naess, Arne. 1973. “The Shallow and the Deep, Long-Range Ecology Movement: A Summary.” *Inquiry* 16 (1–4): 95–100.
54. National Institute of Standards and Technology (NIST). 2023. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. Gaithersburg, MD: NIST. https://doi.org/10.6028/NIST.AI.100-1.
55. Niemiec, Ryan M. 2019. *Character Strengths Interventions: A Field Guide for Practitioners*. hogrefe. ISBN: 9780889374928.
56. Nissenbaum, Helen. 2010. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford, CA: Stanford University Press.
57. Nowak, Martin A. 2006. “Five Rules for the Evolution of Cooperation.” *Science* 314 (5805): 1560–1563. https://doi.org/10.1126/science.1133755.
58. Nussbaum, Martha C. 2011. *Creating Capabilities: The Human Development Approach*. Cambridge, MA: Belknap Press of Harvard University Press.
59. Organisation for Economic Co-operation and Development (OECD). 2019. *Recommendation of the Council on Artificial Intelligence*. OECD/LEGAL/0449. Paris: OECD.

CC BY–NC 4.0

60. Ostrom, Elinor. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press.
61. Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Oxford University Press.
62. Parfit, Derek. 2011. *On What Matters*. Oxford: Oxford University Press.
63. Raji, Inioluwa Deborah, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT '20)*, 33–44. New York: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372873.
64. Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.
65. Regan, Tom. 1983. *The Case for Animal Rights*. Berkeley: University of California Press.
66. Rest, James R. 1986. *Moral Development: Advances in Research and Theory*. New York: Praeger.
67. Roberts, Brent W., Jing Luo, Daniel A. Briley, Peter I. Chow, Rong Su, and Patrick L. Hill. 2017. "A Systematic Review of Personality Trait Change through Intervention." *Psychological Bulletin* 143 (2): 117–141. https://doi.org/10.1037/bul0000088.
68. Roberts, Brent W., Kate E. Walton, and Wolfgang Viechtbauer. 2006. "Patterns of Mean-Level Change in Personality Traits across the Life Course: A Meta-Analysis of Longitudinal Studies." *Psychological Bulletin* 132 (1): 1–25. https://doi.org/10.1037/0033-2909.132.1.1.
69. Ross, Lee. 1977. "The Intuitive Psychologist and His Shortcomings: Distortions in the Attribution Process." In *Advances in Experimental Social Psychology*, vol. 10, edited by Leonard Berkowitz, 173–220. New York: Academic Press.
70. Rothbart, Mary K. 2011. *Becoming Who We Are: Temperament and Personality in Development*. New York: Guilford Press.
71. Rubel, Alan, Clinton Castro, and Adam Pham. 2019. "Agency Laundering and Information Technologies." *Ethical Theory and Moral Practice* 22 (4): 1017–1041. https://doi.org/10.1007/s10677-019-10030-w.
72. Sartre, Jean-Paul. 1946. *Existentialism Is a Humanism*. Paris: Nagel.
73. Scanlon, T. M. 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.
74. Schwartz, Barry, and Kenneth Sharpe. 2010. *Practical Wisdom: The Right Way to Do the Right Thing*. New York: Riverhead Books.
75. Sen, Amartya. 2009. *The Idea of Justice*. Cambridge, MA: Harvard University Press.

CC BY–NC 4.0

76. Shonkoff, Jack P., Andrew S. Garner, et al. 2012. “The Lifelong Effects of Early Childhood Adversity and Toxic Stress.” *Pediatrics* 129 (1): e232–e246. https://doi.org/10.1542/peds.2011-2663.
77. Singer, Peter. 1975. *Animal Liberation*. HarperCollins. ISBN: 978-0-06-171130-5.
78. Simon, Herbert A. 1957. *Models of Man: Social and Rational; Mathematical Essays on Rational Human Behavior in a Social Setting*. New York: John Wiley & Sons.
79. Taiz, Lincoln, et al. 2019. “Plants Neither Possess nor Require Consciousness.” *Trends in Plant Science* 24 (8): 677–687. https://doi.org/10.1016/j.tplants.2019.05.008.
80. Taylor, Paul W. 1986. *Respect for Nature: A Theory of Environmental Ethics*. Princeton, NJ: Princeton University Press.
81. Tett, Robert P., and Denise D. Burnett. 2003. “A Personality Trait-Based Interactionist Model of Job Performance.” *Journal of Applied Psychology* 88 (3): 500–517. https://doi.org/10.1037/0021-9010.88.3.500.
82. Thaler, Richard H., and Cass R. Sunstein. 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven, CT: Yale University Press.
83. Thompson, Dennis F. 1980. “Moral Responsibility of Public Officials: The Problem of Many Hands.” *American Political Science Review* 74 (4): 905–916. https://doi.org/10.2307/1954312.
84. Trist, Eric L., and Ken W. Bamforth. 1951. “Some Social and Psychological Consequences of the Longwall Method of Coal-Getting.” *Human Relations* 4 (1): 3–38. https://doi.org/10.1177/001872675100400101.
85. Trivers, Robert L. 1971. “The Evolution of Reciprocal Altruism.” *Quarterly Review of Biology* 46 (1): 35–57. https://doi.org/10.1086/406755.
86. Tronto, Joan C. 1993. *Moral Boundaries: A Political Argument for an Ethic of Care*. New York: Routledge.
87. United Nations. 1984. *Convention against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment*. New York: United Nations.
88. United Nations Economic and Social Council. 1984. *The Siracusa Principles on the Limitation and Derogation Provisions in the International Covenant on Civil and Political Rights*. New York: United Nations.
89. Weick, Karl E., and Kathleen M. Sutcliffe. 2015. *Managing the Unexpected: Sustained Performance in a Complex World*. 3rd ed. Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/9781119175834.
90. Williams, Bernard. 1981. *Moral Luck: Philosophical Papers 1973–1980*. Cambridge: Cambridge University Press.

CC BY–NC 4.0

91. Wood, Wendy, and David T. Neal. 2007. “A New Look at Habits and the Habit–Goal Interface.” *Psychological Review* 114 (4): 843–863. https://doi.org/10.1037/0033-295X.114.4.843.
92. Young, Iris Marion. 2011. *Responsibility for Justice*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195392388.001.0001.
93. Zimmerman, Michael J. 2020. “Willful Ignorance and Moral Responsibility.” In *Oxford Studies in Normative Ethics*, vol. 10, edited by Mark Timmons. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198867944.003.0004.

CC BY–NC 4.0

2 Operationalization

Part II translates the Potentialism Framework (PF) from conceptual orientation into practical use under real constraints. It proposes a set of lightweight artifacts—compatibility judgments, context snapshots, responsibility maps, options sets, will-regulation prompts, and decision records—that help make ethical reasoning explicit, revisable, and proportional to stakes. It introduces a fast compatibility checklist for time-pressured situations, and a modular decision protocol for higher-impact, contested, or hard-to-reverse choices, with a focus on separating what is known from what is assumed and widening the space of viable options. Part II also clarifies how PF scales between fast and slow modes, and why forced urgency can become a distinct compatibility risk. Finally, it operationalizes "regulated responsibility" (awareness × power) for organizational settings, offering role-and-boundary mapping patterns and a training pathway from novice to practitioner to auditor—aimed at sustaining answerability and protecting dignity of awareness without turning PF into a rulebook or compliance ritual.

CC BY–NC 4.0

2.1 Practical use: what PF outputs

- **In scope:** what PF can *produce* in practice (artifacts that make judgment more legible, revisable, and accountable); how these artifacts relate.
- **Out of scope:** a mandatory procedure; a universal scoring system; anything that replaces moral agency.
- **Notes:** these are optional scaffolds; depth should scale with stakes and real constraints.

PF v2.0 is a **compass**: it helps people and teams orient within messy situations where values, constraints, power, and uncertainty collide. The "outputs" described here are practical artifacts PF **can** produce when its concepts are applied—often partially, depending on time and stakes.

These artifacts are **templates** meant to make reasoning more explicit and revisable, not to replace judgment or moral agency. In some settings, templates can become "the process." PF's intent is the opposite: adapt, simplify, and reorder as needed, and treat any template as *optional scaffolding*, not a compliance ritual.

One practical use of these artifacts is **separation**: keeping conceptual/normative judgment distinct from empirical assumptions. Where an output depends on empirical or technical claims, that layer should be cited or explicitly marked *[NEEDS CITATION]*.

2.1.1 What "outputs" means here

PF does not produce single final answers. It can help generate:

- **Structured judgments** (how compatible is an expression in this context, and why?)
- **Traceable reasoning** (what was considered, what was uncertain, what tradeoffs were accepted?)
- **Role clarity** (who carries which responsibility, given awareness and power?)
- **Actionable next steps** (what to try, avoid, or monitor)

These can be created by individuals, groups, or (in some settings) partially supported by tools—without changing the underlying ethical posture.

2.1.2 Core output artifacts

A simple way to see how these pieces fit:

- **Context snapshot** → clarifies what conditions shape meaning/effects
- **Options set** → makes alternatives and tradeoffs visible

CC BY–NC 4.0

- **Compatibility judgment** → states reasons across the three lenses
- **Responsibility map** → locates where awareness and impact concentrate
- **Decision record** → preserves the above for review, learning, and revisability

Not every situation needs every artifact. PF's recurring question is: *what is the smallest structure that keeps the ethical "contact points" real in this context?*

2.1.2.1 Context Snapshot

Because **context** is part of PF's architecture (not a footnote), a compact "context snapshot" captures conditions most likely to alter compatibility.

Useful dimensions (choose what is salient; note what is unknown):

- relational field (who is involved, what dynamics exist)
- power/impact (who can affect outcomes, at what scale; reversibility)
- resources/constraints (time, money, access, capacity)
- incentives (what is rewarded/punished)
- norms/rules/instructions (explicit and implicit)
- awareness (who can understand likely consequences well enough to steer or contest them)
- internal states (pressures, fears, fatigue, urgency)

The goal is not exhaustiveness but reducing "context collapse"—where decisions are justified abstractly while real drivers stay invisible.

2.1.2.2 Options Set

PF often produces not one "best" expression but a **set of plausible expressions**, each with:

- likely benefits and harms (including dignity risks)
- who bears the costs
- what safeguards or constraints could shift compatibility
- what would need to remain true for acceptability

This reduces false dilemmas and makes tradeoffs visible—without pretending tradeoffs are automatically justified.

2.1.2.3 Compatibility Judgment

A compatibility judgment is a bounded evaluation of **candidate expressions in context** relative to their foreseeable effects on:

CC BY–NC 4.0

- **avoidable suffering/harm**
- **dignity of awareness**
- **others' freedom to regulate their potentials within real constraints (to a reasonable degree)**

This judgment does not "solve" the hard parts of these lenses (e.g., disagreement about what is avoidable, or what counts as a genuine constraint). Its value is practical: it keeps the core questions in view, makes assumptions legible, and makes it easier to revise when context or understanding changes.

It is often worth writing down when stakes are high, uncertainty is meaningful, or a decision may set precedent. In lower-stakes settings, a few bullets or a brief shared agreement can often be enough.

A useful judgment is rarely absolute. It typically reads:

- "**More compatible than** option B under these assumptions…"
- "**Conditionally compatible** if these safeguards hold…"
- "**High risk to dignity** if this power asymmetry is real…"
- "**Unclear** because awareness/impact is uncertain…"

What matters is the *shape of reasoning*: what was treated as foreseeable, what counted as avoidable, where dignity limits were binding, and where uncertainty remained.

Clarifier (to reduce confusion later): a **compatibility judgment** is the *reasons + evaluation*; a **decision record** (see Section 2.1.2.6) is the *container* that preserves the judgment, context, options, and revisit conditions for later review.

2.1.2.4 Will Practice Prompt

Because **Will** is the trained capacity to regulate **how, when, and how intensely** potentials are expressed, PF outputs sometimes include reflective prompts that can help translate ethical insight into a change in expression. Examples include:

- "What potential is driving this expression?"
- "What intensity would meet the purpose with less harm?"
- "What pause or check would reduce avoidable suffering?"
- "If I had to explain this to the most affected person, what would I want to be able to say?"

These are examples, not a checklist. In some moments, a single prompt—or none at all—may be the realistic fit.

CC BY–NC 4.0

2.1.2.5 Responsibility Map

PF links **responsibility** to **awareness and power (impact)**—including the role of Will and ethical insight in regulating expression and answering for effects. A "responsibility map" is an attempt to surface the main actors, affected parties, and influence paths relevant to a decision—recognizing that in complex systems this will be partial and interpretive.

A map may include:

- likely decision-makers and authorizers (sometimes distributed)
- affected parties (including those with low power)
- intermediaries (operators, managers, designers, institutions)
- where awareness is high/low
- where impact is high/low
- where **regulated responsibility (Awareness × Power)** suggests higher ethical weight (as an interaction, not a literal calculation)

This map is not meant to assign blame or create false precision. It is meant to clarify who is positioned to notice, to change course, and to answer for effects—and where responsibility is being implicitly offloaded.

2.1.2.6 Decision Record

When a decision is likely to matter later (stakes, precedent, irreversibility), a brief record **typically captures**:

- the chosen expression (or deferred choice)
- context snapshot summary
- compatibility judgment summary (reasons across harm / dignity / agency)
- key uncertainties and assumptions
- safeguards/monitoring plan (if any)
- revisit trigger ("reassess if X changes")

This supports **answerability**: making it possible to learn, revise, and own effects over time. Like any template, it can be used performatively; its value depends on whether assumptions, tradeoffs, and accountability remain real rather than cosmetic.

2.1.3 How these outputs are used in practice

Rather than a fixed procedure, PF can be applied as a set of **possible moves**. Depending on the setting, you might:

- clarify what expression is being considered (or already happening)

CC BY–NC 4.0

- note a few realistic alternatives
- capture the context features most likely to change outcomes
- sketch a responsibility map (where awareness and impact concentrate)
- write a compatibility judgment that names harms, dignity limits, and agency impacts
- adjust timing, scope, and intensity through will
- record enough to revisit when conditions change

In many settings, only a subset is feasible. PF does not require completeness; it encourages detail proportionate to stakes **to a reasonable degree**—and it treats "stakes" as perspective-sensitive (what is low-stakes to a decision-maker may be high-stakes to those affected, especially under power asymmetry).

Different settings emphasize different artifacts:

- **Individuals** may lean on will prompts plus a lightweight compatibility note.
- **Teams** may emphasize responsibility maps plus decision records.
- **Institutions** may formalize context snapshots and revisit triggers.
- **Tool-assisted workflows** may generate drafts (e.g., an LLM drafting a compatibility scan), but accountability typically remains with people and institutions that authorize deployment and accept impact.

2.1.4 What PF does *not* aim to output

To keep the compass posture intact, PF does not aim to produce:

- a universal rulebook settling moral disputes
- a single metric collapsing dignity, harm, and freedom into one score
- a guarantee of "ethical correctness"
- legal compliance determinations (it can inform them, but is not the same thing)
- a substitute for moral agency or responsibility

PF outputs are intended to be usable under uncertainty and revisable when awareness or power shifts.

2.1.5 Bridge to the next section

Many situations need a faster pass than a full set of artifacts. The next section provides a compact **compatibility checklist** for quick triage when time, attention, or coordination bandwidth is limited.

CC BY–NC 4.0

2.2 Compatibility checklist

- **In scope:** a compact prompt set for fast compatibility triage; designed for readability and adaptation.
- **Out of scope:** a full decision protocol; a scoring system; a "box to tick" that substitutes for judgment.
- **Notes:** treat as prompts, not procedure; if stakes/irreversibility/dispute are high, a fuller pass may fit better.

In PF, when an action is **high-impact** and its effects are **reasonably foreseeable** (given available awareness), responsibility generally scales with awareness and power (impact). That is often reason to **slow down where feasible**—enough to notice avoidable suffering/harm, dignity limits, and who loses freedom-to-regulate under real constraints.

However, real contexts sometimes force speed—crises, deadlines, cascading risks, coordination breakdowns. When quick decisions happen anyway, this checklist is a prompt for remembrance: even under pressure, try to widen the option space, surface key uncertainties, and consider more than the first available path.

This section offers a **compact prompt set** for quick compatibility triage—fast, readable, non-legalistic. It is best treated as **adaptable prompts**, not a required procedure: different situations call for different questions, and the order can be flexible.

2.2.1 If you have ~1 minute

In urgent situations, even a brief scan can surface "pause" signals. This is not a substitute for deeper analysis when time allows.

For a quick orientation, you might consider one prompt from each category:

- **Harm:** What avoidable suffering/harm could this create (given realistic options), and who bears it?
- **Dignity:** Does this plausibly approach **humiliation or destruction of awareness beyond limits**?
- **Agency:** Does this restrict others' freedom-to-regulate beyond what real constraints plausibly justify?
- **Reversibility:** If we're wrong, how hard is this to undo—and can we choose a more reversible variant?
- **Responsibility:** Where do awareness and impact concentrate, and who is positioned to change course?

CC BY–NC 4.0

If any answer feels like “maybe, and we’re rushing,” that can be a signal worth slowing down for, if possible—or escalating to the Decision Protocol.

2.2.2 Quick scan: name the expression and context

If you can, make the object of judgment concrete:

- Identify the specific **expression**, and the few context dimensions most likely to matter (power/impact, incentives, constraints, norms, urgency).
- Note what is uncertain but important—and who is affected if you’re wrong.

2.2.3 The three-lens compatibility prompts

Use these as triangulation. You don’t need certainty—just enough clarity to see what you’re choosing to risk, and which assumptions you’re leaning on.

2.2.3.1 Avoidable suffering / harm

- What harms are foreseeable given what is already known here (even if probabilities are unclear)?
- Which harms look avoidable through changes in scope, timing, intensity, safeguards, or alternatives—given the option-set that is actually available?
- Who bears costs—especially those with low power to refuse, contest, or exit?

2.2.3.2 Dignity of awareness (ceiling check)

A “ceiling” lens: some routes may be unacceptable even if other benefits are large.

- Does this risk **humiliation or destruction of awareness beyond limits** (or require treating that possibility lightly)?
- Are we treating a being/system with awareness as mere instrument—especially under coercion, captivity, dependency, or severe asymmetry?
- If this involves restraint, manipulation, or invasive control: is a less dignity-threatening option being seriously considered?

If you cannot answer these with integrity under time pressure, that itself can be a cue to slow down or escalate.

2.2.3.3 Others’ freedom to regulate potentials (within real constraints, to a reasonable degree)

- Whose freedom to regulate their potentials is reduced—and how?

CC BY–NC 4.0

- Which limits are genuine constraints (resources, safety, coordination), and which reflect design choices or incentives rather than necessity?
- Are we narrowing options temporarily and revisably, or locking them in?

2.2.4 Context prompts (the "why this might go wrong" layer)

If you have a bit more time, it can help to name one or two context realities that often distort judgment:

- **Power/impact:** who can steer outcomes, and who bears costs without voice?
- **Incentives/pressure:** what rewards speed or silence?
- **Awareness gaps:** who can understand consequences, and who carries risk without that visibility?

2.2.5 Regulated responsibility prompts (awareness × power)

A short scan for responsibility drift:

- Who has the most understanding of likely effects, and who has the most practical impact?
- Where might responsibility drift toward low-power parties ("they clicked agree," "they signed the form")?

2.2.6 Options prompts (avoid false dilemmas)

Even two alternatives can change the ethical shape of a choice.

- What is the **least-harm variant** that still serves the core purpose?
- What is a **delay / pause** option that buys learning, consent, or coordination?
- What is a **smaller-scope pilot** (lower impact, higher revisability)?
- What would we do if optimizing for "reduce avoidable suffering/harm" *and* "protect dignity" instead of speed?

2.2.7 Will prompts (regulating intensity, timing, scope)

These prompts can help translate ethical insight into a change in expression (without assuming perfect self-regulation):

- What intensity would meet the purpose with less harm?
- What boundary, pause, or check would reduce foreseeable harm under uncertainty?
- If we had to stand behind this later, what would we change now (scope, timing, reversibility, monitoring)?

CC BY–NC 4.0

Use these as examples, not a completeness test.

2.2.8 Minimal decision note (optional)

If the choice is likely to matter later, even a brief note can support revisability and answerability (where feasible). Useful elements might include:

- what was chosen (the expression)
- why (the key tradeoff)
- what might be wrong (top uncertainties)
- what would trigger reassessment

This is not meant to create routine bureaucracy; it is one possible aid for learning under changing context.

2.2.9 Common failure modes

Checklists can become box-ticking exercises. Their value depends on sincere engagement with the underlying concerns, not mechanical completion.

Quick self-check:

- Are we using these prompts to clarify, or to justify a decision already made?
- Are the most affected parties' interests visible—or only the decision-makers' constraints?
- Are we calling something a "constraint" that is actually a preference or convenience?

If the checklist becomes a tool for post-hoc justification, it stops being useful.

2.2.10 Bridge to the next section

This checklist is designed for fast orientation. When the situation is higher-stakes, strongly contested, difficult to reverse, or dignity-sensitive—or when key answers depend on uncertain assumptions—PF benefits from a more explicit reasoning flow. The next section provides a decision protocol for slowing down in a structured way **without turning PF into a rulebook**.

CC BY–NC 4.0

2.3 Decision protocol

- **In scope:** a repeatable scaffold for higher-stakes or contested cases; separating knowns from assumptions; widening options; making responsibility/power and dignity risks explicit.
- **Out of scope:** legal compliance; formal risk management standards; institutional governance requirements.
- **Notes:** modular and adaptable; the goal is to widen attention under pressure, not to mandate a fixed sequence.

Section 2.2 offers a fast triage. This section is for situations that are **higher-stakes**, **contested**, **hard to reverse**, or **dignity-sensitive**—where a quick scan may not be enough, and where a major risk is locking in a harmful path under pressure, power asymmetry, or uncertainty.

This is a *protocol* in the sense of a **repeatable scaffold**, not a compulsory procedure. In practice, teams can adapt the order, skip parts that don't fit, begin at whichever "move" is most urgent, or loop back as new information arrives. The aim is to widen attention when speed, authority, or conflict would otherwise narrow it.

A common sequence (often iterated) is:

clarify the expression (Section 2.3.3.1)
↓
capture context (Section 2.3.3.2)
↓
map responsibility/power (Section 2.3.3.3)
↓
separate knowns vs assumptions (Section 2.3.3.4)
↓
widen options (Section 2.3.3.5)
↓
evaluate Compatibility (harm / dignity / agency) (Section 2.3.3.6)
↓
work with conflict (Section 2.3.3.7)
↓
regulate expression with Will (Section 2.3.3.8)
↓
set safeguards/revisit triggers (Section 2.3.3.9)
↓
verify answerability (Section 2.3.3.10)

CC BY–NC 4.0

2.3.1 Entry check: are you in “protocol territory”?

A fuller pass **may** be useful when one or more of these is true:

- Impact is large, hard to reverse, or likely to set precedent.
- There is meaningful disagreement about harm, dignity, or fairness.
- Power is concentrated (few can decide; many are affected).
- The choice depends on uncertain empirical assumptions.
- “Urgency pressure” is compressing the option space.

If none applies, the checklist in Section 2.2 may suffice. If you’re unsure, a partial pass—especially **responsibility/power + options**—can still reduce avoidable blind spots.

2.3.2 Compass check in crisis (when time collapses)

Some contexts are genuinely tragic: every available option carries serious harm, and deliberation capacity is constrained. In those cases, the protocol may compress into a few anchors:

- Keep the dignity ceiling in view (humiliation or destruction of awareness beyond limits).
- Prefer the most reversible path available, even if imperfect.
- Make key assumptions and tradeoffs explicit (even as brief bullets), so the decision remains revisable and answerable.
- Leave a trace for revisiting (what happened, why, and what would trigger reassessment).

This does not “solve” tragedy. It aims to preserve ethical contact under pressure.

2.3.3 Protocol moves (modular, often revisited)

2.3.3.1 Clarify the expression and the decision boundary

It often helps to make the object of judgment concrete:

- What **expression** is being considered?
- What is the decision boundary (what is “in scope” now, what is not)?
- What is the time horizon (one-time act, repeated practice, precedent-setting policy)?

A common failure mode in contested situations is debating abstractions while the real expression remains vague.

CC BY–NC 4.0

2.3.3.2 Build a context snapshot (including contested interpretations)

Because context shapes Compatibility, it can help to capture conditions most likely to change consequences:

- power/impact (scale, reach, reversibility)
- incentives (what is rewarded/punished)
- constraints (resources, time, coordination limits)
- norms/rules/instructions (formal and informal)
- awareness (who can understand likely consequences well enough to steer or contest them)
- internal states (fear, fatigue, urgency, anger)

Where interpretations differ—e.g., "this is a constraint" vs "this is a choice"—it can help to mark that explicitly as a point of disagreement rather than assuming a neutral snapshot.

2.3.3.3 Responsibility and power analysis (make asymmetry visible)

In higher-stakes situations, "who decides" can be a hidden driver. A responsibility/power map can focus attention on where ethical weight is concentrating:

- Who has **power/impact** over outcomes (authority, resources, access, leverage, veto points)?
- Who has the most **awareness** of likely effects (and what parts are they blind to)?
- Who is most affected but least able to refuse, contest, exit, or be heard?
- Where might responsibility drift toward low-power parties ("they consented," "they clicked agree," "they could leave")?

This is not about blame. It is about locating where **regulated responsibility—the interaction between understanding and practical impact—concentrates**, so responsibility is less likely to be quietly offloaded onto those least able to carry it.

2.3.3.4 Separate what you *know* from what you *assume*

Before evaluating Compatibility, it often helps to distinguish:

- what is directly known or observed,
- what is inferred,
- what is uncertain but decision-relevant,
- what would change the decision if it turned out false.

This can clarify whether the disagreement is primarily about facts, assumptions, values, or authority (or an overlap).

CC BY–NC 4.0

2.3.3.5 Generate a small options set (widen the possibility space)

When people feel trapped in a false dilemma, ethical reasoning often collapses into "yes/no" under pressure. If feasible, it can help to surface a small set of real alternatives, such as:

- a least-harm version of the proposed expression (reduced scope/intensity),
- a delay/pause option that buys learning, consent, or coordination,
- a smaller-scope pilot with higher revisability,
- the most reversible path available (even if imperfect),
- a different route to the same goal that shifts burdens or preserves more freedom-to-regulate within real constraints.

This is not abstract optimization. It is producing concrete alternatives that can actually be compared.

2.3.3.6 Evaluate Compatibility across the three lenses

For each option, evaluate Compatibility relative to foreseeable effects on the three lenses of Compatibility:

harm reduction, **dignity ceiling**, and **agency preservation**.

This is usually most useful as a **structured judgment with reasons**, not as a binary label.

A. **Avoidable suffering / harm**
    I. What harms are foreseeable given what is known?
    II. Which harms look avoidable through changes in scope, timing, intensity, safeguards, or alternatives—given the option-set surfaced above?
    I. Who bears the costs, especially those with low power?
B. **Dignity of awareness (ceiling check)**
    I. Does this option plausibly approach or cross a dignity ceiling (humiliation or destruction of awareness beyond limits)?
    II. Are we treating a system or being with awareness as mere instrument under coercion, dependency, captivity, or severe asymmetry?
    III. If this involves invasive control, is a less dignity-threatening option being seriously considered?
C. **Others' freedom to regulate potentials (within real constraints, to a reasonable degree)**
    I. Whose freedom-to-regulate is reduced, and how?

CC BY–NC 4.0

II. Which limits appear to be genuine constraints, and which reflect design choices or defaults rather than necessity?
III. Are restrictions temporary and revisable, or do they lock in diminished agency?

Where "real constraints" are disputed, it can help to name the dispute directly: "If constraint X is binding, option A looks more compatible; if X is negotiable/designable, option B may preserve more freedom-to-regulate."

2.3.3.7 Handle conflicts explicitly (without promising resolution)

In contested situations, disagreement can carry information. It can help to explore its main dimensions, even when they overlap:

- **Facts disagreement:** What is disputed? What evidence would change minds?
- **Values disagreement:** Which lens is being weighted differently (harm vs dignity vs freedom-to-regulate), and why?
- **Authority disagreement:** Who legitimately decides, and on what basis, given awareness and power?

If low-power affected parties are not meaningfully represented, your picture may be incomplete. Where direct participation isn't feasible, one partial substitute is to assign an explicit role to argue from the standpoint of those most affected and least empowered.

If conflict persists, outcomes that often preserve ethical room include: choosing a more reversible option, reducing scope/intensity, adding safeguards and revisit triggers, and recording dissent rather than forcing false consensus.

2.3.3.8 Choose with Will (regulate expression, not only direction)

Once an option is selected, PF emphasizes **regulating expression**:

- Can intensity be reduced while meeting the purpose?
- Can timing change (pause, stage-gate, wait for critical information)?
- Can scope be narrowed (who/where/how long)?
- What boundary, check, or safeguard would reduce avoidable harm under uncertainty?

This is where ethical insight becomes operational: not only "what we choose," but "how we express it."

CC BY–NC 4.0

2.3.3.9 Safeguards, monitoring, and revisit triggers

For higher-impact choices, it is often helpful to clarify:

- monitoring targets (signals that harm/dignity risk is rising),
- who has authority to slow/stop/adjust,
- what would trigger reassessment (context change, new evidence, credible complaint, threshold crossed).

Safeguards can reduce risk, but they can also be used to soothe doubt. Keeping the compatibility judgment explicit helps prevent "monitoring" from becoming a substitute for ethical evaluation.

2.3.3.10 Closing verification: answerability and learning

As a closing check, it can help to revisit the responsibility/power picture from earlier:

- Who is answerable for effects as they unfold?
- Are the most affected parties visible in the reasoning, not only as abstract "stakeholders"?
- If this goes wrong, will the system learn—or will costs be externalized and forgotten?

This is not a guarantee of correctness. It is a commitment to remain ethically engaged as context shifts.

2.3.4 Bridge to the next section

This protocol tends to produce artifacts—compatibility judgments, context snapshots, responsibility maps, options sets, and safeguard/revisit triggers—that benefit from consistent recording when decisions matter later. The next section defines a **decision record standard** to keep judgments transparent, reviewable, and less prone to post-hoc rationalization.

CC BY–NC 4.0

2.4 Decision record standard

- **In scope:** what to document so judgments are reconstructable, reviewable, and revisable; how to keep assumptions and uncertainty visible.
- **Out of scope:** legal compliance; formal risk management; mandated retention rules.
- **Notes:** a "shared shape," not bureaucracy; record at the safest useful level of specificity.

PF treats decisions—especially high-impact ones—as **expressions in context**. A decision record is a lightweight way to keep the expression, context, and reasoning visible over time: what you saw, what you assumed, what tradeoffs you accepted, and what would justify revisiting.

This is a **standard** in the sense of a shared shape for recording decisions—not a bureaucratic requirement. The aim is proportionality: in lower-stakes settings, a few lines may be enough; in higher-stakes contexts, more structure can help the decision stay reviewable, revisable, and answerable.

A caution: documentation does not "prove" compatibility. A record can be used well (to support learning and correction) or misused (as performative compliance, blame-shifting, or moral cover). Treat the record as a constructed account shaped by incentives and pressures—so it should surface key assumptions and uncertainties a future reader would need to question.

2.4.1 When a decision record tends to add the most value

A record tends to be most valuable when one or more of these is true:

- impact is large, dignity-sensitive, or difficult to reverse
- the decision sets precedent (a pattern likely to repeat)
- uncertainty is significant, and being wrong would be costly
- power is concentrated, and those affected have limited voice or exit
- disagreement exists (facts, values, or authority), even if unresolved
- safeguards or monitoring are part of the rationale for proceeding

The point is not completeness. It is supporting reconstruction of the reasoning—so later review is possible without relying only on memory, status, or hindsight.

CC BY–NC 4.0

2.4.2 Three scalable “shapes” (illustrative)

Choose a level of detail that fits the situation and constraints. These are examples, not tiers to satisfy.

A. **Pocket note (2–6 lines)**
   When time is limited but you want traceability.
   I. Expression (what was decided)
   II. Context/Why (main driver + tradeoff, 1–2 sentences)
   III. Top uncertainty (1 sentence)
   IV. Revisit trigger (what would reopen it)

B. **Structured record (often useful for higher-impact choices)**
   A compact structure that can survive disagreement.
   I. Expression + boundary + time horizon
   II. Context snapshot
   III. Options considered + reversibility notes
   IV. Compatibility summary (three lenses, as reasons)
   V. Regulated responsibility (Awareness × Power) map
   VI. Safeguards + monitoring targets + revisit triggers
   VII. Uncertainties + dissent/representation gaps (if present)
   VIII. Basis for key claims (conceptual / mixed / evidence-leaning)

C. **Extended record (when stakes, power, or dignity risks are extreme)**
   Adds deeper context and stronger auditability.
   I. fuller assumptions register (what would flip the decision)
   II. explicit representation of low-power affected parties’ standpoint
   III. stronger rationale for irreversibility (if any)
   IV. clearer repair/learning plan if harms occur

2.4.3 Decision record prompts (adapt as needed)

These prompts are modular. Use what fits; omit what doesn’t; add what’s missing. Where documenting details creates privacy/safety risk, record at the safest useful level of specificity.

Basics

- Short label (so it can be found later)
- Date/time
- Decision-maker(s) (individual/team/role)
- **Basis for key claims:**

CC BY–NC 4.0

- *conceptual* (primarily normative/interpretive judgment)
- *mixed* (concept + some evidence)
- *evidence-leaning* (primarily external data/analysis)

(This is not a PF-specific taxonomy; it is an epistemic cue for what kind of scrutiny the claim invites—not a status claim.)

Expression + decision boundary

- Expression (the concrete action/output/policy being chosen)
- Decision boundary (what is in scope now, and what is not)
- Time horizon (one-time act, repeated practice, precedent-setting policy)

Context snapshot (the few variables that could flip the judgment)

- Power/impact (scale, reach, reversibility)
- Incentives (what is rewarded/punished)
- Constraints (resources, time, coordination limits)
- Norms/rules/instructions (formal and informal pressures)
- Awareness: who can understand likely consequences well enough to steer or contest them—and who bears risk without that visibility
- Internal states (urgency, fatigue, fear, anger—if they materially shape expression)
- *Record-pressure note (optional):* what incentives or constraints might be shaping the record itself (e.g., fear of blame, speed pressure, status dynamics)

Options considered + reversibility

- Chosen option
- Least-harm / reduced-intensity variant
- Pause/delay or smaller-scope pilot
- Reversibility notes: what is reversible, what is not, and why irreversibility is a risk *or* a necessary feature for safety in this context

If only one option was feasible, note why (constraints, timing, coordination failure).

Key assumptions and uncertainties

- Known/observed
- Inferred
- Uncertain but decision-relevant
- What would flip the decision (evidence/assumptions that, if changed, would alter compatibility)

CC BY–NC 4.0

Compatibility summary (across the three lenses)

- Avoidable suffering / harm: foreseeable harms; what was judged avoidable given the option-set available here
- Dignity of awareness (ceiling check): any dignity ceiling concerns; why the chosen option is judged to stay within dignity limits (or what dignity-cost is being taken under constraint)
- Others' freedom to regulate potentials (within real constraints, to a reasonable degree): whose freedom-to-regulate is reduced; which limits are constraints vs. design/defaults/incentives; whether restrictions are temporary/revisable or lock in diminished agency

Regulated responsibility (Awareness × Power) map

- who had the most awareness of likely effects
- who had the most power/impact to steer outcomes
- who bears the costs with low ability to refuse/exit/contest
- who can still change course if harms emerge
- who is answerable if the outcome is worse than expected

Will-regulation note (how expression is being regulated)

- intensity / timing / scope adjustments made (or not made), and why
- what boundary, check, or safeguard reduces avoidable harm under uncertainty

Safeguards, monitoring targets, and revisit triggers

- Safeguards (what reduces foreseeable harm under uncertainty)
- Monitoring targets (what signals you intend to track)
- Authority to slow/stop/adjust (who can act "in time to matter")
- Revisit triggers (what threshold, complaint, evidence, or context-change reopens the decision)

Where monitoring is part of the rationale for proceeding, it helps to keep it tied to the compatibility reasoning above—so "we're monitoring" does not become a substitute for ethical judgment.

Dissent, unresolved conflict, and representation gaps (if present)

- Facts dispute
- Values tradeoff
- Authority dispute
- Who was not represented (especially low-power affected parties)

CC BY–NC 4.0

- How that gap was handled

Closing: answerability and repair orientation

- what you would do if harms appear (repair / narrowing scope / stopping conditions)
- what you would change next time (a short learning note, if appropriate)

2.4.4 Bridge to the next section

Decision records scale with time and pressure: sometimes a pocket note is all that is feasible; sometimes the situation justifies a slower, fuller record. The next section explains that scaling explicitly—how PF shifts between **fast mode** and **slow mode**, and why forced urgency is itself a compatibility risk.

CC BY–NC 4.0

2.5 Fast mode and slow mode

- **In scope:** why PF distinguishes fast and slow modes; why forced urgency is a compatibility risk; how to create "room" for slow mode when stakes justify it.
- **Out of scope:** a theory of two "systems" in the mind; mandated organizational timelines; formal incident management.
- **Notes:** operational labels for feasibility under constraints; mode choice is a form of regulating deliberation.

PF is meant to work under real constraints. Many decisions are expressions in context—and contexts differ in time pressure, uncertainty, incentives, and the distribution of awareness and power. The same ethical insight can be expressed in two broad modes:

- **Fast mode:** action under constraint (time, attention, coordination), using a small set of salient cues and safeguards.
- **Slow mode:** deliberate evaluation, with enough space to widen options, surface assumptions, and make tradeoffs explicit.

These are not personality types or a claim about two fixed "systems" in the mind. They are operational labels for what becomes feasible under different constraints. PF borrows the shorthand from decision-making and naturalistic judgment research without claiming a full cognitive model. *[Kahneman 2011; Klein 1998]*

Mode choice itself is an expression of Will in a practical sense: it is part of regulating *how* deliberation happens, *when* it happens, and *how intensely* the system commits attention before acting.

2.5.1 Fast mode

Fast mode is often unavoidable: emergencies, cascading risks, thin information, coordination breakdowns. Under those constraints, attention narrows and the option space can collapse to "do something now." *[Kahneman 2011]*

A PF-oriented fast mode often emphasizes:

- **Stabilize harm trajectories:** reduce intensity, pause escalation, prevent foreseeable avoidable harms.
- **Honor dignity limits as a ceiling:** avoid routes that plausibly cross dignity boundaries, even when the situation is messy.
- **Make one or two regulation moves:** adjust timing or intensity of the expression, or change how it is expressed (narrower channel, tighter safeguards), rather than trying to "solve" the full situation at once.

CC BY–NC 4.0

- **Leave a trail for revisiting:** even a brief note (what was done, why, top uncertainty, revisit trigger) can support later answerability—especially when incentives later reward confident stories.

A fast-mode expression can sometimes be judged compatible within severe constraints if it stabilizes acute harm and stays inside dignity limits. Its ethical weight is often in what it enables next (recovery, correction, de-escalation), not in comprehensive foresight.

Fast mode can also be a common site for rationalization. The risk is not speed itself; it is speed that hides power, suppresses uncertainty, or locks in irreversibility without acknowledging the cost.

2.5.2 Slow mode

Slow mode becomes relevant when stakes are higher, impacts harder to reverse, dignity risks are central, or disagreement persists. It is where PF's architecture becomes more fully expressible: naming the expression, capturing key context variables, widening options (including more reversible variants where feasible), and evaluating Compatibility across the three lenses—avoidable suffering/harm, dignity of awareness, and freedom-to-regulate.

In practice, a PF-oriented slow-mode evaluation often explores:

- clarifying what is being decided (expression + boundary),
- distinguishing what is known from what is assumed,
- generating a small set of realistic alternatives (including a more reversible path, where feasible),
- making responsibility/power concentration explicit (regulated responsibility: Awareness × Power),
- setting safeguards and revisit triggers,
- and producing a decision record so the reasoning remains reviewable and revisable.

Slow mode does not guarantee correctness. It creates more opportunity for harms, dignity limits, and responsibility concentration to become visible enough to be regulated rather than ignored.

2.5.3 Forced urgency as a compatibility risk

PF treats forced urgency as a distinctive risk pattern: urgency that is not only "real," but sometimes produced or amplified by incentives, authority, or avoidable delays—so scrutiny becomes impractical and responsibility can drift.

CC BY–NC 4.0

A practical marker is when time pressure is intensified in ways that predictably reduce meaningful review—for example: last-minute requirements, schedules that prevent consultation, incentives that reward speed over careful evaluation, or decision windows set to expire before dissent can surface.

Forced urgency can be a compatibility risk because it can:

- compress options into false dilemmas ("only one path"),
- privilege perspectives of high-power actors who can act quickly,
- reduce meaningful representation of those most affected,
- increase the chance of locking in a harmful path under uncertainty.

This is not a claim that every urgent situation is engineered. It is an orientation: when urgency appears, it can help to ask whether it is purely constraint-driven—or partly incentive-driven.

2.5.4 Creating room for slow mode (when stakes justify it)

In many systems, slow mode is not simply "chosen"; it is enabled. Enabling moves can be costly and can be misused as performative compliance—so the compass aim is proportionality: create enough room for key concerns and tradeoffs to surface, without turning "process" into a substitute for judgment.

Illustrative patterns (where feasible) include:

- **A pause lever:** a right-sized way to slow escalation (time-boxed pause, stage gate, "stop-the-line" authority).
- **Protected dissent / bad-news flow:** channels that allow concerns to surface without punishment. *[Weick and Sutcliffe 2007]*
- **Role design choices:** distinguishing proposal, evaluation, and answerability roles can reduce collapse into justification in some contexts *[Nemeth et al. 1990; Tetlock 1985]*, but can also fragment responsibility and add coordination cost *[Becker and Murphy 1992; Thompson 1967]*.
- **Pre-committed revisit triggers:** making it legitimate to change course when thresholds are crossed or new evidence arrives.
- **Early avoidance of lock-in:** delaying irreversible commitments until core uncertainties narrow, when possible.

These are examples, not a minimum spec. The point is to notice which structural features in a context are expanding or shrinking the practical possibility of ethical scrutiny.

CC BY–NC 4.0

2.5.5 Switching modes

PF treats mode choice as revisable. A decision can begin in fast mode and move to slow—or the reverse—based on context changes.

Common signals to consider shifting **fast → slow** include:

- new information increases likely impact,
- new stakeholders become clearly affected,
- evidence of harm or dignity risk rises,
- disagreement persists and begins to shape outcomes,
- you recognize an upcoming lock-in point (a commitment that significantly reduces future reversibility).

Common signals to consider shifting **slow → fast** include:

- credible time windows close,
- delay plausibly increases avoidable harm,
- coordination breaks down and a stabilizing move becomes necessary—while keeping the dignity ceiling in view.

The compass aim is not "always slow." It is staying sensitive to when more structure is ethically warranted—and when speed is necessary but ethically costly.

2.5.6 Bridge to the next section

Mode choice interacts with roles and power: who gets time, who gets heard, and who can slow or steer outcomes. The next section shows how to map **regulated responsibility (Awareness × Power)** across individuals, teams, executives, and system boundaries—so ethical weight is less likely to drift onto low-power parties.

CC BY–NC 4.0

2.6 Roles and responsibility mapping in organizations

- **In scope:** mapping regulated responsibility across roles/teams/executives/system boundaries; common mismatch patterns; how mapping supports answerability and dignity-sensitive decisions.
- **Out of scope:** a prescriptive org chart; a complete governance standard; legal liability allocation.
- **Notes:** the map is itself an expression in context; treat it as revisable and contestable.

PF treats organizational choices as expressions in context—often expressed through structure: who is authorized to act, who can question, who can stop escalation, and who bears downstream costs.

A responsibility map is one way to make those structures legible. It is not a hunt for blame. It is a way to check whether the distribution of responsibility—obligation proportional to awareness and power (impact) to use will in line with ethical insight, and to answer for effects—corresponds to how decisions and impacts actually flow, especially when systems are complex, distributed, and partly automated.

A key caution: the map itself is also an expression in context. What gets included, excluded, or "made official" can reflect incentives, status pressures, and power. The practical aim is not neutrality; it is to make assumptions explicit enough to be questioned, corrected, and owned.

2.6.1 Why map roles (PF view)

When outcomes go well, responsibility often feels obvious. When outcomes go poorly, organizations can fall into familiar failure modes: responsibility diffusion ("everyone touched it, no one owned it"), or the reverse—scapegoating ("someone must be punished, therefore they must have been responsible"). A PF map aims at a third posture: answerability proportional to awareness and power to use will in line with ethical insight.

Conceptually, this proportionality matters because it helps prevent two recurring distortions:

- **Blind power:** high impact without sufficient understanding of consequences.
- **Burden dumping:** harms or cleanup work landing on low-power parties without meaningful voice, exit, or recourse.

Both can erode others' freedom to regulate their potentials within real constraints (to a reasonable degree).

CC BY–NC 4.0

2.6.2 What to map

A responsibility map is often most illuminating when it captures only the dimensions that actually shape the decision and its effects. Depending on context, it can help to note:

- **The expression:** what is being decided or done (policy, deployment, enforcement, escalation, rollback, etc.).
- **System boundary:** where the "system" ends for this decision (team, org, vendors, contractors, external tools, third-party AI modules, users).
- **Roles and authorities:** who can propose, evaluate, approve, deploy, monitor, throttle/rollback, and stop escalation.
- **Awareness distribution:** who can understand likely consequences well enough to steer or contest them (and who bears risk without that visibility).
- **Power/impact distribution:** who can materially shape outcomes (scale, access, resources, coercive capacity, irreversibility).
- **Incentives and pressures:** what is rewarded or punished (speed, growth, compliance, silence), and what that does to scrutiny and escalation.
- **Dignity sensitivity:** where humiliation, coercion, dependency, captivity, or severe asymmetry might plausibly arise.
- **Answerability & repair:** who is positioned to explain choices, respond to harms, and change course "in time to matter."

2.6.3 Mapping prompts (adapt as needed)

These prompts are modular. Use what fits; omit what doesn't; add what's missing. Because a map can become "the story people tell," it can help to capture uncertainties and contested points—not only a clean narrative.

1. **Name the decision and boundary**
    a. Decision / expression:
    b. Boundary: what is in scope (and what is not)?
    c. Time horizon: one-time act, repeated practice, precedent-setting policy?
2. **List actors and roles**
    For each actor/role (individual, team, vendor, tool, AI component), note:
    a. Role in the chain: propose / evaluate / approve / deploy / operate / monitor / stop authority / affected party
    b. What they can actually do: practical authority, not just nominal responsibility
3. **Sketch Awareness and Power (regulated responsibility)**
    Treat this as a heuristic about relative levels, not a precise measurement. Also note that power can shape whose knowledge "counts" as awareness.

CC BY–NC 4.0

   a. High awareness / high power: typically carries high regulated responsibility.
   b. High power / low awareness: **blind power** risk (impact without sufficient understanding).
   c. High awareness / low power: **powerless insight** risk (knowing without ability to steer).
   d. Low awareness / low power: often vulnerable to being used as cover ("we consulted") or burdened without recourse.
4. **Identify likely burden, drift, and incentive pressure**
   a. Who bears costs with limited ability to refuse/exit/contest?
   b. Where could responsibility quietly drift (through handoffs, automation, layered approvals)?
   c. What incentives or deadlines compress scrutiny—especially near irreversible commitments?
5. **Define answerability**
   a. who is expected to explain the choice and its safeguards,
   b. who can change course,
   c. who owns repair if harms occur,
   d. what triggers revisit (thresholds, complaints, evidence, context change).

Where stakes are high, capturing this in a standard decision record can preserve traceability over time—so later narratives don't replace the reasoning that actually governed the choice.

2.6.4 Role design choices (tradeoffs, not defaults)

Mapping often reveals role-design questions. A few patterns recur:

- Themes in organizational research highlight the value of dissent, structured challenge, and clear accountability for resisting premature consensus and improving scrutiny under pressure. *[Nemeth 1986; Schweiger, Sandberg, and Ragan 1986; Tetlock 1985]*
- One way some organizations attempt to operationalize this is by distinguishing proposal, evaluation, and answerability roles. This can reduce collapse into justification in some contexts, but it can also fragment responsibility ("I only proposed; they evaluated") and add coordination cost. *[Becker and Murphy 1992; Thompson 1967]*
- Stop authority and pause levers can protect dignity-sensitive decisions from being forced into fast mode by default—but can also be misused for obstruction or blame-avoidance if not paired with clear answerability.

CC BY–NC 4.0

- Protected channels for bad news (hotlines, escalation paths, anonymous reporting) can help preserve awareness in the system—but their existence alone can become performative. They tend to matter most when they are credible (people can speak without punishment) and connected to someone with power to act. *[NIST 2023; Weick and Sutcliffe 2007]*

These are design considerations, not a minimum spec. PF's interest is the alignment: as awareness and power rise, responsibility rises—and the organization's structure should make that legible rather than hide it.

2.6.5 Bridge to the next section

Once roles and regulated responsibility are mapped, the next question is how people learn to use the map without turning PF into a compliance ritual. The next section proposes a training pathway—from novice to practitioner to auditor—so the framework stays a compass under real incentives, time pressure, and organizational politics.

CC BY–NC 4.0

2.7 Training pathway: novice to practitioner to auditor

- **In scope:** a progression of postures for learning PF in practice; what tends to change with training; common drift patterns.
- **Out of scope:** a universal certification; a mandated curriculum; a single "correct" pedagogy.
- **Notes:** these are postures of fluency and review, not new core terms.

PF is meant to be usable under real constraints—time pressure, incentives, politics, and imperfect information. That usability depends less on memorizing "rules" and more on training skill: noticing context, regulating expression through Will, and applying Ethics as practiced judgment.

This section offers a training pathway—not a credential ladder and not a compliance regime. It describes three postures people may occupy (sometimes informally, sometimes formally): **novice**, **practitioner**, and **auditor**. *(These labels describe training milestones and review postures, not additional core conceptual units of PF.)* The point is to make learning legible: what tends to change as someone becomes more capable, and how responsibility may reasonably scale when awareness and real influence (power/impact) increase.

2.7.1 A useful distinction: cultivating capacity vs regulating expression

PF distinguishes Potential from Expression. In training terms, people often benefit from holding two time horizons at once:

- **Cultivation over time:** strengthening what tends to be available when triggered (skills, habits, attention patterns, ways of seeing context).
- **Regulation in the moment:** shaping what is enacted here and now—especially how, when, and how intensely something is expressed.

This is not a strict boundary. The practical point is simpler: when stakes are immediate, the ethical work is often regulating the expression under real constraints—even while longer-horizon cultivation continues where feasible.

2.7.2 Pathway overview (non-linear by design)

Think of these as three postures of fluency, not a ladder:

1. **Novice posture:** can name the architecture and make a basic compatibility-oriented move under constraint.
2. **Practitioner posture:** can apply PF reliably in contested, higher-stakes contexts, and can hold a decision process together without turning it into ritual.

CC BY–NC 4.0

3. **Auditor posture:** can evaluate decisions and systems for responsibility alignment, rationalization risk, and dignity-ceiling pressure—while staying alert to how reviewing can be used as a weapon.

People can move between these postures depending on domain and context. In a new domain, even an expert may return to a novice posture.

2.7.3 Novice posture

Orientation: basic PF literacy + one or two reliable moves, often in fast mode.

Capabilities that often develop here include:

- Spotting the architecture: identify expression vs context; notice obvious incentives and pressures.
- A small Will move: look for one regulation lever (reduce intensity, change timing, change channel) when stakes justify it.
- A compatibility scan: a quick check across harm / dignity / agency.
- Minimal traceability: when a decision is likely to be revisited, leave a small trail for later review.

Possible novice drift patterns include:

- Turning PF into slogans rather than situational judgment.
- Collapsing the three lenses into a single feeling or a single metric.
- Using "constraints" as a blanket excuse without naming which constraint actually binds.

2.7.4 Practitioner posture

Orientation: apply PF under real stakes, disagreement, and organizational friction.

A practitioner can work in both modes: act in fast mode without confusing speed with justification, and create or protect slow mode when warranted.

Capabilities that often develop here include:

- Context scrutiny: distinguish material constraints from design/default choices—and from incentive pressures that compress scrutiny.
- Decision shaping: widen options when possible; when not possible, make limits explicit.
- Assumption discipline: separate what is known, inferred, and uncertain; name what would flip the decision.

CC BY–NC 4.0

- Compatibility reasoning: make tradeoffs explicit across the three lenses, including who loses freedom-to-regulate and whether that loss is constraint-driven or design/default-driven.
- Responsibility mapping: identify where Awareness × Power concentrates, and where it drifts through handoffs or automation.
- Repair orientation: keep "what would repair look like?" active early, not only after harm occurs.
- Documentation without ritual: produce records that preserve reasoning and revisit triggers, without using paperwork to simulate ethics.

Possible practitioner drift patterns include:

- Process substitution: mistaking records, checklists, or meetings for ethical judgment.
- Committee fog: slow mode that diffuses responsibility ("we all agreed") rather than clarifying answerability.
- One-lens capture: letting urgency, dignity, or harm dominate so completely that the others disappear.

2.7.5 Auditor posture

Orientation: evaluate decisions and systems for responsibility alignment and rationalization risk—without becoming a compliance weapon.

"Auditor" here can be an internal or external review posture, not necessarily a formal job title. An auditor is not "more ethical" by title. The value of the posture is trained skepticism about how decisions are represented: where power hides, where incentives distort records, where dignity risks are minimized in language, and where low-power parties carry costs without recourse.

A key PF reminder is that **auditing is also an expression in context**: shaped by the auditor's incentives, independence, access to information, and the organization's appetite for real scrutiny. When those conditions are weak, "audit" can become theater—or an instrument of blame.

Capabilities that often develop here include:

- Narrative skepticism: treat records and maps as constructed expressions, not neutral snapshots. Look for what is omitted, softened, or reframed.
- Responsibility alignment checks: compare formal roles to actual awareness and power; flag major gaps (blind power / powerless insight patterns).

CC BY–NC 4.0

- Dignity ceiling vigilance: notice where language reframes humiliation/coercion/dependency risks as merely "efficient" or "necessary."
- Evidence reasoning (not citation theater): distinguish conceptual/normative claims from empirical/technical ones; scrutinize what is supported, what is assumed, and what would count against it.
- Revisit integrity: assess whether revisit triggers are actionable, owned, and likely to be honored under incentive pressure.
- Audit humility: notice how review can drift toward defensiveness, box-ticking, or blame games—and counter that drift in framing, scope, and follow-through.

Possible auditor drift patterns include:

- Becoming an enforcement surrogate: "audit" used to punish rather than to increase answerability and learning.
- Over-indexing on documentation polish instead of outcome-relevant reasoning quality.
- Treating mapped responsibility as static, ignoring how crises and incentives reshuffle practical power.

2.7.6 Avoiding compliance drift

The drift patterns flagged throughout Part II—box-ticking, process substitution, blame theater—apply here too. Two orienting questions that can help counter ritualization are:

1. In what ways did this artifact make real tradeoffs more visible—and in what ways might it protect status?
2. Does it clarify who can change course "in time to matter"—or does it diffuse responsibility behind process?

These questions do not have automatic answers. Their value is in making disagreement discussable, and in keeping attention on real effects rather than procedural comfort.

2.7.7 Bridge to Part III

Part II focused on practice under constraints: checklists, protocols, records, modes, and responsibility mapping. Part III turns to measurement and validation: how to test whether PF-guided practices are helping (or not) across compatibility outcomes, dignity protections, and responsibility drift—without reducing ethics to a single metric.

CC BY–NC 4.0

2.8 References (Part II)

1. Becker, Gary S., and Kevin M. Murphy. 1992. “The Division of Labor, Coordination Costs, and Knowledge.” *The Quarterly Journal of Economics* 107 (4): 1137–1160. https://doi.org/10.2307/2118383.
2. International Organization for Standardization and International Electrotechnical Commission (ISO/IEC). 2023. *ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system.* Geneva: ISO. https://www.iso.org/standard/81230.html.
3. International Organization for Standardization and International Electrotechnical Commission (ISO/IEC). 2023. *ISO/IEC 23894:2023 Information technology — Artificial intelligence — Guidance on risk management.* Geneva: ISO. https://www.iso.org/standard/77304.html.
4. Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.
5. Klein, Gary. 1998. *Sources of Power: How People Make Decisions*. Cambridge, MA: MIT Press.
6. National Institute of Standards and Technology (NIST). 2023. *Artificial Intelligence Risk Management Framework (AI RMF 1.0).* NIST AI 100-1. Gaithersburg, MD: NIST. https://doi.org/10.6028/NIST.AI.100-1.
7. Nemeth, Charlan J., Ofra Mayseless, Jeffrey W. Sherman, and Yvonne Brown. 1990. “Exposure to Dissent and Recall of Information.” *Journal of Personality and Social Psychology* 58 (3): 429–437. https://doi.org/10.1037/0022-3514.58.3.429.
8. Nemeth, Charlan Jeanne. 1986. “Differential Contributions of Majority and Minority Influence.” *Psychological Review* 93 (1): 23–32. https://doi.org/10.1037/0033-295X.93.1.23.
9. Schweiger, David M., William R. Sandberg, and James W. Ragan. 1986. “Group Approaches for Improving Strategic Decision Making: A Comparative Analysis of Dialectical Inquiry, Devil’s Advocacy, and Consensus.” *Academy of Management Journal* 29 (1): 51–71. https://doi.org/10.2307/255859.
10. Tetlock, Philip E. 1985. “Accountability: The Neglected Social Context of Judgment and Choice.” In *Research in Organizational Behavior*, 297–332. Greenwich, CT: JAI Press.

CC BY–NC 4.0

11. Thompson, James D. 1967. *Organizations in Action: Social Science Bases of Administrative Theory.* New York: McGraw-Hill.
12. Weick, Karl E., and Kathleen M. Sutcliffe. 2007. *Managing the Unexpected: Resilient Performance in an Age of Uncertainty*. 2nd ed. San Francisco: Jossey-Bass.

CC BY–NC 4.0

3 Measurement and validation

Part III develops a practical stance on measuring and validating Compatibility without turning the Potentialism Framework (PF) into a single score or a compliance ritual. It explains why Compatibility is inherently multi-lens—tracking avoidable harm, the dignity of awareness, and the freedom to regulate potentials under real constraints—and why any measurement will therefore be partial and context-sensitive. The Part distinguishes what can be measured directly, what can only be proxied, what requires qualitative judgment, and what remains inherently contested even in good-faith evaluation. It then offers prototype instruments (case snapshots, lens-separated rubrics, and decision records) designed to improve legibility, contestability, and answerability while resisting proxy gaming and audit substitution. Finally, Part III proposes calibration loops, testable practice-trace hypotheses, and a research agenda for piloting, comparing, and revising these tools over time—aiming for disciplined learning rather than definitive ethical "proof."

CC BY–NC 4.0

3.1 The measurement problem

- **In scope:** why "Compatibility" resists a single metric; what kinds of evidence can still be relevant; how to measure without turning PF into a compliance score.
- **Out of scope:** a universal compatibility index; definitive tests for dignity violations; guarantees that a "good score" implies real-world compatibility.
- **Notes:** Measurement supports learning and answerability; it does not replace judgment. When metrics are used, treat them as fallible evidence with known failure modes.

PF's central evaluation lens—**Compatibility**—evaluates an **expression** in **context** relative to its reasonably foreseeable effects on: (a) avoidable suffering/harm, (b) the **dignity of awareness**, and (c) others' freedom to regulate their potentials within real constraints (to a reasonable degree). Because this is a **three-lens** evaluation (harm, dignity, agency), measurement is inherently partial.

This is not a flaw to "fix." It is a design choice: PF stays context-sensitive and ceiling-aware, which limits what measurement can responsibly claim.

PF therefore treats measurement as **purpose-bound**: useful for learning, coordination, and answerability—while resisting the temptation to turn proxies into verdicts. Classic failure modes of proxy governance apply: when a measure becomes a target, it tends to stop being a good measure. *[Goodhart 1975; Campbell 1979; Strathern 1997]*

3.1.1 Why compatibility resists a single metric

Compatibility is not one thing.

1. **Avoidable suffering/harm** often invites quantitative signals (incidents, error rates, complaints, near-misses). But "avoidable" depends on alternatives, constraints, and what was reasonably foreseeable given available awareness. Counting harm does not automatically answer avoidability.
2. **Dignity of awareness** functions as a ceiling principle. Ceilings do not behave like smooth scores: they create **high-salience thresholds** and "stop-and-question" zones—especially under uncertainty about awareness and vulnerability.
3. **Freedom to regulate potentials within real constraints** (agency) is often indirect: exit options, contestability, retaliation risk, dependence, and information asymmetry rarely collapse into one number.

A single "compatibility score" would therefore hide tradeoffs and invite gaming. It can also create a dangerous illusion: that ethics has been *computed* rather than reasoned.

CC BY–NC 4.0

3.1.2 What measurement can still do

Measurement can still be useful—if its role is clear.

- It can **surface patterns** (where harm clusters, where complaints concentrate, where drift occurs).
- It can **support contestability** (making claims checkable rather than purely rhetorical).
- It can **improve answerability** (documenting what was known, what was assumed, what was ignored).
- It can **enable iteration** (revising rubrics and safeguards when reality diverges from expectation).

The key move is to treat metrics as **evidence**, not as permission slips.

3.1.3 What counts as evidence in PF-style evaluation

Because Compatibility is multi-lens, PF benefits from keeping evidence plural. Common evidence types include:

1. **Outcome evidence:** what happened (harms, benefits, downstream effects), including lagging indicators.
2. **Process evidence:** how the decision was made (who was consulted, what uncertainties were recorded, what safeguards were chosen).
3. **Behavior under conditions:** how an agent/system behaves across contexts (including under pressure, novelty, or incentives).
4. **Lived experience / testimony:** how affected parties report impacts, dignity threats, coercion, or constraint—especially where harms are hard to count.

No single category settles Compatibility. Together, they can make judgments more reviewable and less prone to rationalization.

3.1.4 Proxy failure modes (why "scoring" goes wrong)

When metrics are treated as targets, predictable distortions follow:

- **Goodharting / Campbell-style distortion:** optimizing the proxy instead of the underlying aim. *[Goodhart 1975; Campbell 1979]*
- **Audit substitution:** replacing lived reality with paperwork success ("we have the form"). *[Strathern 1997]*
- **Selective visibility:** measuring what is easy to count, not what matters most.

CC BY–NC 4.0

- **Power shielding:** those most affected may have the least ability to generate "official" signals.

These are reasons to **design measurement with humility**, not reasons to abandon it.

3.1.5 A practical stance: "measured, but not settled"

A compatible measurement posture often looks like:

- Make explicit **which lens** a metric bears on (harm? dignity? agency?).
- Name what the metric **cannot** tell you, and what it might incentivize.
- Pair metrics with **qualitative judgment** and **contestability** (who can challenge the account?).
- Keep a minimal **decision record** for high-stakes choices: what was chosen, why, what was uncertain, what would trigger reassessment.

3.1.6 Bridge to the next section

With the measurement problem named, the next step is to draw a boundary: **what can be measured directly, what can be proxied, what needs qualitative judgment, and what remains inherently contested even in good-faith practice.**

CC BY–NC 4.0

3.2 What can be measured and what cannot

- **In scope:** practical boundaries for measurement—what tends to be directly measurable, what can be proxied, what requires qualitative judgment, and what remains inherently contested.
- **Out of scope:** a universal "compatibility test"; definitive dignity-violation thresholds; claims that measurement can replace ethical judgment or responsibility.
- **Notes:** This is an orientation map, not a checklist. The goal is to keep claims honest about what kind of support they provide.

A useful boundary is to distinguish four intertwined aspects of compatibility evaluation:

1. **Direct indicators** (often countable, but still interpreted)
2. **Proxies** (measurable signals with context-dependent meaning)
3. **Qualitative judgments** (reasoning that does not reduce to numbers)
4. **Inherently contested questions** (value-laden thresholds where good-faith disagreement can persist)

PF does not require all four in every case. The point is to keep clear **what kind of claim** is being made, so measurement supports judgment rather than quietly replacing it.

3.2.1 What is often measurable (direct indicators)

Examples that are often measurable with locally clear definitions:

- incident counts, error rates, near-miss rates
- response times, uptime/downtime, recovery times
- complaint volumes and resolution times
- documented policy violations or escalation events
- access logs, override events, audit trails (where applicable)

Direct indicators are still interpreted: their meaning depends on reporting incentives, detection capacity, and who is empowered to register harm.

3.2.2 What is sometimes measurable only as a proxy

Many ethically central features show up indirectly:

- **dignity risk signals:** humiliation patterns, dehumanizing treatment, forced exposure, coercive degradation
- **agency / constraint signals:** exit options, retaliation risk, informed consent quality, dependency structure

CC BY–NC 4.0

- **foreseeability signals:** whether warnings existed, whether decision-makers had access to relevant information

Proxies are useful for *triage* and pattern detection, but they rarely settle the question alone. A key discipline is to treat "no signal" as weak evidence of "no risk," especially where reporting is unsafe or harms are normalized. *[CONCEPTUAL]*

3.2.3 What requires qualitative judgment

Some judgments remain qualitative even with excellent data:

- what counted as "avoidable" given real constraints
- whether mitigation was credible or performative
- whether tradeoffs crossed dignity ceilings
- whether affected parties had meaningful ways to contest, refuse, or exit

Qualitative judgment can still be **rigorous** if it is legible: separate observation from inference; name uncertainties; show how competing reasons were weighed.

3.2.4 What remains inherently contested (even in good faith)

"Inherently contested" does not mean "anything goes." It means evidence stays entangled with thresholds and values. Examples:

- **Dignity ceilings under uncertainty:** what level of destruction/humiliation is "beyond limits" in a concrete high-stakes setting?
- **Agency vs. benefit tradeoffs:** when does "consent" become too constrained (economically, socially, institutionally) to count as meaningful?
- **AI-adjacent contested question (example):** when a system reliably models consequences and adapts its behavior over time, what level of deference or protection is owed under uncertainty about the system's awareness and vulnerability?

These questions can be argued better or worse, but they may not converge to a single, stable metric.

3.2.5 Awareness and measurability (clarified[1])

Awareness, as PF uses the term, is operational: the capacity to understand consequences, model self/others over time, and regulate expression. In principle, these capacities can be

[1] Here „clarified" signals a more careful formulation, not a new doctrine or stronger claim.

CC BY–NC 4.0

probed via behavior and performance under varied conditions. In practice, tests remain fallible, can be gamed, and can miss internal limitations—especially under distribution shift or strategic presentation. *[CONCEPTUAL]*

3.2.6 "Measurement-supporting conditions" (scaffolding, not scores)

When stakes are high, measurement tends to be more useful when paired with supporting conditions such as:

- **legibility:** clear separation of observation / proxy / inference
- **contestability:** credible channels for challenge without retaliation
- **traceability:** who decided, on what basis, with what uncertainties
- **repairability:** mitigation and compensation pathways if harms occur
- **misuse-resistance:** likely gaming routes and incentives named

These conditions are not proofs. They are scaffolding that can reduce self-deception and audit substitution. *[Goodhart 1975; Campbell 1979; Strathern 1997]*

3.2.7 Bridge to the next section

With measurement boundaries clear, the next step is practical: **prototype instruments and rubrics** that respect those boundaries—explicitly partial, testable, and revisable rather than presented as definitive compatibility metrics.

CC BY–NC 4.0

3.3 Prototype instruments and rubrics

- **In scope:** early rubrics and capture tools as prototypes; how to structure evidence and judgment without reducing Compatibility to one score; how prototypes stay revisable.
- **Out of scope:** a mandatory procedure; "one true rubric"; any claim that filling out a template implies ethical adequacy.
- **Notes:** The artifacts below work together: a **reasoning flow** helps evaluators think; a **capture template** helps them record what that thinking produced.

Prototypes are useful when they do two things at once: (1) make evaluation more legible and contestable, and (2) resist becoming a performative substitute for responsibility.

3.3.1 Prototype mindset: why "explicitly provisional" matters

A prototype can drift when it becomes, for example:

- a status symbol ("we did the PF form"),
- a weapon ("your score is low"),
- a shield ("the rubric cleared us"),
- a replacement for lived experience ("the numbers say it's fine").

Naming these risks up front helps keep the tool honest. *[Strathern 1997]*

3.3.2 A compact case snapshot (capture what matters first)

A case snapshot is a short, structured description that makes evaluation possible without pretending to be complete. It often includes:

- **the expression:** what is being done (not only what is intended)
- **the context:** key constraints and drivers (relationships, incentives, norms, resources)
- **affected parties:** who is impacted, especially those with low power
- **time horizon:** immediate effects vs delayed effects
- **uncertainties:** what is unknown, disputed, or inferred

Optional add-on for higher stakes: a one-line note on **regulated responsibility** (awareness × power/impact) to keep answerability visible without turning it into arithmetic.

3.3.3 A non-mandatory reasoning flow (illustrative)

Rather than a fixed procedure, evaluators may find a loose flow helpful. For example:

CC BY–NC 4.0

- Start by naming the **expression** and the **context** features most likely to change meaning or impact.
- Identify affected parties, including those likely to bear costs without voice.
- Gather **outcome**, **process**, **behavior-under-conditions**, and **lived experience** evidence as available—while naming what is missing.
- Evaluate across the three lenses:
    - **harm:** what suffering is likely, and what was avoidable given real constraints?
    - **dignity:** are plausible ceiling risks in play (destruction/humiliation beyond limits)?
    - **agency:** who loses realistic freedom to regulate their potentials, and why?
- State the judgment as a **reasoned account**, not a score: what supports it, what weakens it, what would change it.
- For higher stakes: record who had awareness, who had power/impact, what safeguards exist, and what triggers reassessment.

This is meant to guide thinking, not standardize conscience.

3.3.4 A rubric that captures reasoning (without collapsing to one number)

Usable rubric often separates:

- **claims by lens** (harm / dignity / agency)
- **evidence type** (direct indicator / proxy / qualitative)
- **confidence / uncertainty** (what is robust vs what is conjecture)

One simple capture format is a three-row table (one per lens) with columns for: “main concern,” “supporting evidence,” “counter-evidence,” “key uncertainty,” and “what would change the judgment.”

If teams want a shorthand label, one optional **illustrative prototype** is to tag each lens with:

- **concern level** (low / medium / high), and
- **confidence** (low / medium / high).

This is not a metric; it is a visibility aid. Treat it as revisable and easy to discard if it starts behaving like a score. *[CONCEPTUAL]*

CC BY–NC 4.0

3.3.5 High-stakes dignity ceiling checks (a prototype for “tragic choice” contexts)

Sometimes every available option appears to carry dignity risk. In such cases, a prototype can help evaluators avoid false cleanliness:

- explicitly name the dignity concern and why it is “ceiling-like” here,
- describe what makes options constrained (and which constraints are structural vs chosen),
- identify mitigation and repair pathways (even if partial),
- record dissent and unresolved dispute rather than forcing consensus,
- set a reassessment trigger (what evidence or change would reopen the choice?).

This does not “solve” tragic tradeoffs. It preserves legibility and answerability when resolution is not available.

3.3.6 Bridge to the next section

Once prototypes exist, a practical question follows: **can different evaluators use them and still coordinate enough to learn—without forcing false consensus or erasing context?** That is the focus of calibration and inter-rater reliability.

CC BY–NC 4.0

3.4 Calibration and inter-rater reliability

- **In scope:** how evaluators align enough for practical use; what "reliability" can and cannot mean for PF; how to calibrate while preserving context-sensitivity and dissent.
- **Out of scope:** treating high agreement as proof of ethical adequacy; forcing convergence on contested value questions; "one rubric to rule them all."
- **Notes:** Reliability is a coordination signal. It can reveal whether a tool is usable and learnable, not whether a judgment is morally correct.

Calibration asks whether evaluators can apply shared prototypes in a way that is:

- **legible** (they can explain why they judged as they did),
- **comparable** (differences can be diagnosed),
- **revisable** (the prototype changes when it fails),
- **contestable** (dissent can be recorded without retaliation).

3.4.1 What reliability can mean here

In PF settings, "reliability" is often best treated as:

- agreement on **what is being claimed** (observation vs proxy vs inference),
- agreement on **what evidence bears on which lens**,
- clearer understanding of **why** disagreement exists when it persists.

High agreement can still be wrong. Low agreement can still be honest—especially in inherently contested zones.

3.4.2 A practical calibration loop (illustrative)

A simple calibration loop might look like:

1. Choose a small set of cases with varied stakes and contexts.
2. Have evaluators independently produce short lens-separated accounts.
3. Compare: where do differences come from—facts, proxies, thresholds, values, or missing context?
4. Revise the prototype: clarify terms, add prompts, remove misleading fields.
5. Repeat with new cases.

This is partly technical and partly interpretive: revising prototypes changes what becomes visible. *[CONCEPTUAL]*

CC BY–NC 4.0

3.4.3 Protecting calibration from power distortion (minimal safeguards)

Calibration can be distorted when:

- senior voices set the “correct” answer early,
- dissent becomes costly,
- the goal shifts from learning to passing an audit.

A minimal countermeasure is to preserve structured dissent: record minority views, uncertainties, and the reasons disagreement persisted.

3.4.4 Bridge to the next section

Once calibration is possible, we can ask a further question: **what changes in practice when PF is used—changes that can be checked in pilots, retrospectives, or case comparisons without pretending compatibility is fully measurable?** That is the purpose of testable implications.

CC BY–NC 4.0

3.5 Testable implications

- **In scope:** translating PF into hypotheses and checkable patterns in pilots, audits, and retrospectives—especially about practice traces (legibility, contestability, answerability, follow-through).
- **Out of scope:** claiming PF is "proven"; promising outcome gains; treating any single metric (including agreement rates) as proof of ethical adequacy; reducing the three lenses to one score.
- **Notes:** Many PF claims are conceptual/normative. "Testable implications" here means: which parts of practice become more legible, more contestable, and more revisable when PF artifacts are used?

PF-relevant features leave traces that can be checked without collapsing to a compatibility score:

- **artifact traces:** what appears (or is missing) in case snapshots, decision records, post-mortems
- **reasoning traces:** whether observation/proxy/inference are separated; whether uncertainties are named
- **contestability traces:** whether affected parties can challenge the account without retaliation
- **follow-through traces:** whether reassessment triggers and repair pathways are actually used

These are not moral proofs. They are checkable signs of whether a framework changes practice in the way it claims to.

3.5.1 Candidate hypotheses (practice traces)

1. **H1 — Lens separation improves diagnostic clarity.**
   When evaluators separate harm/dignity/agency in records, disagreements may become easier to localize (e.g., disagreement about dignity ceilings vs disagreement about avoidability). This can be checked by comparing how often disagreements are "diagnosed" into categories rather than treated as vague conflict.
2. **H2 — Proxy humility reduces audit substitution.**
   When prototypes explicitly label proxies and likely gaming routes, teams may be less likely to treat proxy improvement as ethical clearance (measurable as fewer "score-only" justifications in reviews). *[Goodhart 1975; Campbell 1979; Strathern 1997]*

CC BY–NC 4.0

3. **H3 — Contestability increases earlier in the lifecycle.**
   In settings using PF artifacts, stakeholder challenge or dissent may appear earlier (before deployment / decision finalization) rather than only after harm occurs.
4. **H4 — Dignity ceiling risks become more "noticed in time" (when plausible).**
   Where dignity threats are plausible, explicit ceiling prompts may increase the frequency of "stop-and-question" moments (e.g., escalations, refusal considerations, redesign) relative to comparable cases without such prompts.
5. **H5 — Agency/constraint mapping changes what counts as "consent."**
   When constraint and retaliation risks are recorded, some cases previously treated as consensual may be reclassified as constrained or coercive (measurable via changed justifications and mitigation choices).
6. **H6 — Calibration reduces accidental drift (without forcing consensus).**
   Repeated calibration cycles may increase consistency in what evidence is treated as bearing on which lens, while still preserving documented disagreement in contested zones.
7. **H7 — Over-standardization tradeoff pattern (candidate failure mode).**
   Increasing standardization can raise agreement while lowering sensitivity to context and minority harm signals—especially in low-power populations. Treat as a candidate risk pattern to watch for.
8. **H8 — Responsibility mapping shifts post-hoc attribution patterns.**
   When records explicitly note who had awareness and power/impact, blame may shift away from low-power operators and toward upstream decision structures (measurable in post-mortems).
9. **H9 — "Agency laundering" becomes easier to detect.**
   If procedures diffuse answerability by treating formal participation as real agency, PF-style records may make that diffusion more legible. ("Agency laundering": procedural formalities that obscure where awareness and power/impact actually lie, thereby weakening answerability.)

3.5.2 Bridge to the next section

If these hypotheses define what could be checked, the next step is methodological: **how do we design studies and learning loops that can check practice traces, resist self-confirmation, and improve the prototypes over time?**

CC BY–NC 4.0

3.6 Research agenda and study designs

- **In scope:** a research agenda for evaluating and improving PF over time—what to study first, what evidence to collect, and study designs that check practice traces without pretending to quantify compatibility.
- **Out of scope:** claiming PF is validated (or superior); prescribing one "correct" methodology; using research outputs to override dignity ceilings or responsibility.
- **Notes:** The goal is answerable learning loops: what changes when PF is used, what fails, what improves with revision—and for whom?

A pilot can be "successful" even without proving moral truth, if it shows improvements like:

- **legibility:** clearer separation of observation / proxy / inference; uncertainties stated
- **contestability:** credible challenge channels; dissent recorded without penalty
- **answerability:** clearer mapping of awareness and power/impact in decisions
- **follow-through:** reassessment triggers used; mitigation/repair activated when reality diverges
- **prototype revision:** templates change in response to failures (not ritualized)

These are measurable as practice traces (document analysis, interviews, review outcomes), not as moral clearance.

3.6.1 A staged research sequence (start small, then scale)

A practical sequence often looks like:

1. **Artifact uptake studies:** do people actually use the prototypes, and where do they break?
2. **Calibration studies:** do evaluators converge on how to *apply* the templates (and diagnose disagreement)?
3. **Comparative case studies:** compare similar cases with/without PF artifacts for legibility, contestability, follow-through.
4. **Longitudinal drift studies:** does the prototype become ritualized, weaponized, or gamed over time? *[Strathern 1997]*
5. **High-stakes "stress tests":** examine whether dignity/agency concerns are noticed earlier under pressure (without assuming success).

3.6.2 Study design options (examples)

- **Before/after within an organization:** compare decisions pre-PF vs post-PF on legibility/contestability/follow-through.

CC BY–NC 4.0

- **Matched case comparisons:** pair cases with similar stakes/context; compare evaluation quality traces.
- **Retrospective audits:** analyze post-mortems for whether harms were foreseeable, whether dissent existed, whether repair happened.
- **Field experiments on prompts/templates:** test alternative prototype designs for clarity and gaming resistance (small scope, ethical safeguards).

3.6.3 Case selection and "safe-case bias" (defined)

A recurring risk is **safe-case bias**: selecting cases that are already low-conflict, well-documented, or low in power asymmetry—making the framework look cleaner than it is under real strain. To reduce this risk, intentionally include cases with:

- high power/impact and uneven awareness,
- contested stakeholder accounts,
- delayed or diffuse harms,
- incentives for performative compliance.

3.6.4 Evidence portfolios: assembled accounts, not neutral containers

A portfolio of "evidence" is always curated: what gets included, excluded, and emphasized is shaped by incentives and power. A useful safeguard is to require portfolios to state:

- what was **not** measured and why,
- whose voices are missing,
- what would count as disconfirming evidence.

This keeps the learning loop honest without pretending neutrality. *[CONCEPTUAL]*

3.6.5 Revision criteria (keep prototypes alive)

A prototype warrants revision when, for example:

- it increases paperwork without improving contestability or follow-through,
- it becomes a score that substitutes for responsibility,
- it consistently misses harms to low-power groups,
- it produces agreement by flattening context (see the over-standardization risk).

Revisions should be tracked: what changed, why, and what failure mode it addressed.

3.6.6 Bridge to the next section

Part III focused on measurement and validation without collapsing Compatibility into a single metric. The next part turns to a harder realism: **how frameworks are misused, how**

CC BY–NC 4.0

power distorts ethical tools, and how to design for misuse-resistance and contestability in the first place.

CC BY–NC 4.0

3.7 References (Part III)

1. AI Safety Atlas. 2025. *AI Safety Atlas*. Markov Grey and Charbel-Raphaël Segerie. PDF report.
2. Campbell, Donald T. 1979. “Assessing the Impact of Planned Social Change.” *Evaluation and Program Planning* 2 (1): 67–90. [https://doi.org/10.1016/0149-7189(79)90048-X](https://doi.org/10.1016/0149-7189%2879%2990048-X).
3. Goodhart, Charles. 1975. “Problems of Monetary Management: The U.K. Experience.” *Papers in Monetary Economics* 1975/1: 1–20. Sydney: Reserve Bank of Australia.
4. Strathern, Marilyn. 1997. “‘Improving Ratings’: Audit in the British University System.” *European Review* 5 (3): 305–321. [https://doi.org/10.1002/(SICI)1234-981X(199707)5:3%3C305::AID-EURO184%3E3.0.CO;2-4](https://doi.org/10.1002/%28SICI%291234-981X%28199707%295:3%3C305::AID-EURO184%3E3.0.CO;2-4).

CC BY–NC 4.0

4 Misuse-resistance and power realism

Part IV examines how Potentialism Framework (PF) is intended to function under real-world pressure: when incentives, institutions, and uneven power shape which expressions are possible and which stories get told about them. It names recurring failure modes—Compatibility-washing, metric substitution, and institutional capture—while offering a misuse-resistant posture grounded in contestability, Responsibility scaling, and the **Dignity of awareness** ceiling. It also clarifies how structural constraints can narrow agency without automatically excusing harm, why baseline dignity thresholds constrain "contextual justification," and when PF needs supplementation by domain evidence, technical safety methods, and enforceable governance. The aim is not a rulebook, but a sturdier compass: one that stays oriented when convenience, defensibility, or power would otherwise bend it.

CC BY–NC 4.0

4.1 Misuse-resistance goals and design posture

- **In scope:** how PF thinking resists misuse, rationalization, and "safe-sounding" drift; what a misuse-resistant posture looks like in design, policy, and practice.
- **Out of scope:** exhaustive threat modeling, security engineering details, or legal compliance frameworks.
- **Notes:** This section is conceptual: it aims to improve *how we look* for misuse, not to guarantee prevention.

PF is meant to function in the real world, where power, incentives, and narrative pressure shape what gets built and justified. A "Compatibility-aligned" label can become a tool of persuasion, and a dignified framing can be used as cover for degrading practice. A misuse-resistant posture treats those risks as part of the ethical terrain, not as an external add-on.

4.1.1 Core misuse risks PF should anticipate

PF misuse often comes in forms that look reasonable at first glance:

4.1.1.1 Legitimation drift

Language shifts faster than practice. An organization adopts PF vocabulary, then uses it to legitimize what it would have done anyway—especially when the vocabulary reduces scrutiny rather than increasing it.

4.1.1.2 "Declared intention" substitution

Good intentions are treated as sufficient evidence of ethical alignment. PF's orientation is different: intentions matter, but *constraints, impacts, and lived experience* remain part of the evaluation.

4.1.1.3 Metric substitution

A narrow proxy becomes the whole story. This is attractive because it is legible and defensible, but it can erase dignity-ceiling risks or redefine harms as "out of scope."

4.1.1.4 Dignity as ornament

Dignity language becomes ceremonial rather than binding. The misuse pattern is not "talking about dignity," but *treating dignity-ceiling risk as rhetorically important and operationally optional*.

CC BY–NC 4.0

4.1.2 A misuse-resistant design posture

A PF-aligned posture tends to treat misuse-resistance as a *design constraint on reasoning*, not a public-facing claim:

- **Prefer reversible moves** where stakes are high and uncertainty is real.
- **Keep contestability alive** (real channels to challenge decisions, not just documentation).
- **Bind rhetoric to procedures:** if dignity is invoked, there should be visible consequences for how decisions are made and reviewed.
- **Scale Responsibility with power/impact:** the more power and awareness an actor has, the less credible "we didn't know" becomes in predictable harm patterns.

4.1.3 Practical warning signs

Misuse risk rises when you see combinations like:

- "We are aligned with PF" paired with low transparency about tradeoffs and failures.
- A shrinking set of admissible evidence ("only the dashboard counts").
- Dignity-ceiling concerns routed into PR language rather than operational escalation.
- Accountability dispersed so that no one can change the outcome even when harm is recognized.

4.1.4 Questions that keep the posture honest

- What would change *if* this were incompatible—procedurally, materially, institutionally?
- Who can trigger reconsideration, and what happens when they do?
- What is being optimized, and what is being silently treated as acceptable collateral?
- Where are we vulnerable to "safe-sounding" rationalization?

4.1.5 Bridge to the next section

Misuse-resistance becomes concrete when we name specific drift modes. The next section focuses on one of the most common: **Compatibility-washing**, where Compatibility language is used to neutralize scrutiny while coercion and degradation proceed under a "benevolent" frame.

CC BY–NC 4.0

4.2 Compatibility-washing

- **In scope:** what Compatibility-washing is, how it appears, and how to detect it without turning PF into a purity test.
- **Out of scope:** adjudicating all contested moral language, or deciding which institutions are "truly aligned."
- **Notes:** This section treats washing as a *risk pattern*—often incentive-shaped—rather than only bad faith.

Compatibility-washing occurs when Compatibility language is used to legitimize an expression while avoiding the substantive questions PF is meant to keep visible: real constraints, lived impact, Responsibility scaling, and dignity-ceiling risk. Washing can be deliberate; it can also be a sincere rationalization shaped by incentives.

4.2.1 How Compatibility-washing happens

4.2.1.1 Scope-squeezing

The "Compatibility evaluation" is defined so narrowly that the most important harms are declared irrelevant.

4.2.1.2 Substitution-by-proxy

A convenient proxy (a metric, a certification, a checklist) is treated as if it *is* Compatibility.

4.2.1.3 Narrative capture

The story of the action becomes more protected than the actual people affected by it. Dissent is reframed as confusion, hostility, or ignorance.

4.2.1.4 Moral credentialing

Past "goodness" is treated as evidence that present actions are compatible—reducing scrutiny exactly when scrutiny is needed.

4.2.2 Operational warning signs

- Compatibility claims are strong while tradeoffs are vague.
- "Harm reduction" language appears without clarity about *whose harm*, *whose agency*, and *what coercion costs*.
- Contestability exists on paper but is predictably ineffectual.
- Responsibility is described as "shared" while power/impact is concentrated.

CC BY–NC 4.0

- **Awareness-as-status.** Differences in awareness are recast as differences in human worth, so that elites, experts, or more articulate actors are treated as more "dignified," while lower-status people are treated as ethically lesser or as fit to be overruled for their own good.

4.2.3 Belief systems and sacred framing

Compatibility-washing can also appear through sacred or identity-protective frames: "This is holy," "This is who we are," "Questioning this is violence." PF does not try to dissolve belief systems. It asks whether sacred language is being used to exempt a power arrangement from contestability and from dignity-ceiling constraints.

4.2.4 Example tension: parents, religion, and a child's developing freedom

This example is illustrative rather than dispositive. It is meant to show how Compatibility language can be used (sometimes sincerely) to rationalize coercion.

A parent believes strict religious practice prevents "moral harm," and frames enforcement as care: *harm reduction*, *protection*, *for their own good*. The child experiences the same enforcement as humiliation, fear, and loss of practical freedom to regulate their own emerging potentials (within real constraints, to a reasonable degree). The conflict is not only about beliefs; it is also about power/impact and who bears the cost of the "protective" intervention.

PF's role is not to settle theology or parenting doctrine. It is to keep several questions visible at once:

- **What is being treated as "harm," and whose experience counts as evidence?**
- **What forms of constraint are being used, and what do they do to dignity and agency over time?**
- **How is Responsibility distributed?** Protective intent does not automatically offset loss of dignity, avoidable suffering, or freedom-to-regulate; developmental limitation does not erase dignity—it changes what guidance is proportionate and what safeguards are appropriate.

Compatibility-washing appears when "care" language is used to shortcut these questions—treating declared benefit as if it were a Compatibility argument, while the child's actual constraints, suffering, and degradation are treated as acceptable collateral.

CC BY–NC 4.0

4.2.5 Bridge to the next section

Washing becomes durable when it is reinforced by institutions—procedures, incentives, and accountability structures. The next section addresses **institutional capture**: how the surrounding system can adopt PF language while reshaping it into a shield.

CC BY–NC 4.0

4.3 Institutional capture

- **In scope:** how PF language and processes get captured; why contestability matters; what institutional patterns tend to protect dignity rather than merely perform it.
- **Out of scope:** diagnosing motives, naming villains, or providing a universal governance model.
- **Notes:** Capture can be "normal-looking." The signal is the gap between language and operational constraint.

"Institutional capture" here points to a familiar failure mode: PF language is adopted, but the adoption primarily changes *how things are described*—not the incentives, constraints, and accountable choices that shape expressions in practice. From the inside, capture does not always feel like a conspiracy. It can feel like "normal procedure," especially when incentives reward smoothness, defensibility, and plausible deniability.

One recurring diagnostic is **contestability**: do affected parties (or credible advocates) have practical ways to challenge a decision, with a realistic possibility of triggering reconsideration, mitigation, repair, or at least a recorded rationale for declining?

4.3.1 Common capture modes

These patterns can co-occur. They are framed as failure modes of institutional use, not as moral verdicts on people.

As illustrations: a "PF review" that happens only after a decision is effectively locked; a grievance channel that exists on paper but predictably retaliates against complainants; or a metric dashboard that becomes the sole admissible evidence while lived experience is treated as anecdote.

1. **Procedural substitution:** "We followed the process" replaces substantive ethical evaluation.
2. **Metric substitution:** A narrow proxy becomes the only truth the system recognizes.
3. **Responsibility diffusion:** Accountability is spread thin enough that no one can be held meaningfully responsible, even when power and awareness were concentrated upstream.
4. **Dissent suppression:** The system treats contestation as noise, hostility, or noncompliance rather than as a key input to ethics.

CC BY–NC 4.0

4.3.2 Contestability as a protective principle

Contestability is not "people can complain." It is *structured capacity* for challenge and revision. In PF terms, contestability protects against two central drifts:

- **Narrative monopoly:** only the institution's story is legible.
- **Unilateral power:** those most affected have no credible leverage to shape constraints, remedies, or exit.

4.3.3 Design patterns that tend to improve contestability

These are not a checklist. They are overlapping families of supports:

1. **Accessible reasons and traceability:** People can see why a decision was made (to a reasonable degree), what evidence mattered, and which assumptions were decisive.
2. **Independent review capacity:** Review is not structurally dependent on the same incentives that produced the decision.
3. **Protected dissent and anti-retaliation reality:** Channels matter less than whether using them predictably harms the user.
4. **Remedy power (not only "feedback"):** Contestability is weak if challenges cannot trigger mitigation, repair, reversal, or exit options.
5. **Standing for advocates and affected groups:** Institutions can increase ethical resolution by making room for credible advocates—not only insiders.
6. **Treating dignity-ceiling risks as escalation conditions:** Where plausible dignity risks exist, institutions often benefit from slowing down where feasible, increasing contestability, and clarifying who can pause, halt, or modify an expression—and on what grounds. The point is not that "dignity" wins every conflict; it is that the Dignity of awareness ceiling raises the burden of justification and care when destruction or humiliation of awareness is plausibly at stake (see Section 1.9).
7. **Clear responsibility mapping:** When something goes wrong, it should be legible who had power to prevent it, who benefited, and who can repair.

4.3.4 Cultural norms that prevent performative PF

Institutional design matters, but so do norms:

- Treating "unknowns" as a reason to increase humility and review—not to declare safety.
- Rewarding truthfulness about tradeoffs and failures.
- Holding high power/impact actors to higher explanatory and repair burdens.

CC BY–NC 4.0

4.3.5 Bridge to the next section

Institutional capture is one way power reshapes ethics in practice—by reshaping what counts as evidence, compliance, and accountability. The next section widens the lens to **structural constraints on will**: how coercion, scarcity, incentives, trauma, and institutional design shape which expressions are realistically available, and how Responsibility is evaluated under those constraints.

CC BY–NC 4.0

4.4 Structural constraints on will

- **In scope:** how real-world constraints shape feasible expressions; how PF evaluates Responsibility under constraint without collapsing into excuse-making or blame.
- **Out of scope:** clinical claims about trauma or universal causal models of behavior.
- **Notes:** The aim is ethical realism: Responsibility that tracks power/impact, awareness, and feasible choice.

PF talks about Will as a trained capacity, but trained capacities operate inside conditions: coercion, deprivation, information control, and institutional role constraints. Ethical evaluation that ignores constraints often becomes moralizing; ethical evaluation that treats constraints as total exculpation often becomes permissive. PF is trying to keep a workable middle.

4.4.1 Common constraint families

- **Material constraints:** Poverty, dependency, disability, and scarcity can narrow the option set so severely that "choice" becomes mostly formal.
- **Coercion and threat:** When refusal is punished (physically, economically, socially), autonomy claims should be treated cautiously.
- **Informational constraints:** If people are denied reasons, misled about stakes, or prevented from seeing alternatives, Responsibility evaluation shifts toward those shaping the information environment.
- **Institutional role constraints:** Roles and hierarchies can constrain what someone is permitted to do, say, or refuse—especially when discretion stays upward while accountability and penalties are pushed downward. In such contexts, "individual Will" may be present, but the option set is institutionally fenced. PF's Responsibility lens tends to point back toward those who design, enforce, or benefit from the fence: they usually carry more obligation to widen feasible options, reduce coercion, and make challenge-and-appeal routes workable rather than merely nominal.
- **Time constraints and crisis framing:** Urgency can be real, but "crisis" can also become a tool for bypassing contestability and normal safeguards.

4.4.2 Why this matters ethically (Responsibility scaling)

Constraints change what it is reasonable to expect. PF's Responsibility view scales with awareness and power/impact: the more someone can shape constraints and foresee outcomes, the more obligation they tend to carry to prevent dignity-ceiling harms and to provide repair paths.

CC BY–NC 4.0

4.4.3 Avoiding moralizing and avoiding permissiveness

A constraint-aware posture avoids two common errors:

- **Moralizing error:** treating constrained behavior as if unconstrained, then blaming the person nearest the harm.
- **Permissive error:** treating constraints as erasing all agency, then letting high power/impact actors off the hook by declaring "complexity."

4.4.4 Practical posture under constraint

In general, PF-aligned practice tends to ask:

- What options were actually available (not theoretically available)?
- How was refusal priced?
- Who could widen the option set at the lowest cost?
- What repair or exit routes exist for those bearing the harms?

4.4.5 Relation to dignity-ceiling risks

Constraints matter especially near dignity-ceiling risks: coercion plus degradation can create a trap where the harmed person is also blamed for not escaping. PF treats this as a Responsibility red flag: it often indicates upstream power is being insulated while downstream agency is being moralized.

4.4.6 Bridge to the next section

Taking constraints seriously does not imply that "anything goes." The next section names baseline constraints and dignity thresholds anchored in the **Dignity of awareness** ceiling: limits on what contextual justification can reasonably excuse, especially where destruction or humiliation of awareness is plausibly at stake.

CC BY–NC 4.0

4.5 Baseline constraints and dignity thresholds

- **In scope:** baseline constraints that function as strong safeguards; how they interact with contextual justification; how they shape Responsibility claims.
- **Out of scope:** absolute rule sets that "solve" ethical conflict.
- **Notes:** These are "high-threshold" constraints—meant to resist rationalization, not to eliminate hard cases.

Part IV deals with real-world misuse pressures: washing, capture, and constrained agency. Baselines matter because they serve as a counterweight to the most common rationalization move: "the context made it necessary" (especially when dignity-ceiling harms are being normalized).

4.5.1 Baselines versus contextual justification

Baselines are not meant as a simplistic list of prohibitions. They are meant as *pressure-tests*:

- If an action plausibly destroys or humiliates awareness, the burden of justification rises sharply (see Section 1.9).
- If contextual justification is offered, the justification should address real alternatives, proportionality, and repair—not only intent.

4.5.2 Five recurring dignity-threshold constraints

These are patterns where PF tends to treat the ethical bar as high, and where "context" often becomes a laundering channel if not carefully handled:

A) **Avoidable coercion over the core of a person's agency** (especially when "care" language is used as cover).
B) **Systematic humiliation or degradation** as a tool of control or compliance.
C) **Destruction, disabling, or instrumentalization of awareness** when plausible alternatives exist.
D) **Non-consensual extraction or manipulation** that prevents meaningful refusal or informed participation.
E) **Severe harm with predictable downstream denial** (harms routed through systems that prevent repair and accountability).

Where uncertainty is high and stakes are severe, PF tends to favor precautionary postures (slowing down where feasible, widening contestability, strengthening repair/exit options).

CC BY–NC 4.0

4.5.3 What baselines do (and do not do)

Baselines do not eliminate tradeoffs. They do:

- prevent "context" from becoming an all-purpose excuse for dignity harm;
- highlight where institutional incentives need redesign;
- clarify where Responsibility claims require more than procedural compliance.

4.5.4 Responsibility when baselines are pressured

Baseline breaches tend to show up where power is uneven, option sets are narrow, and narratives are curated. PF's Responsibility scaling matters here (see Section 1.11):

- When high power/impact actors design or authorize systems that predictably route dignity-ceiling risks downstream while remaining insulated from the worst effects, PF's Responsibility lens is meant to surface that mismatch.
- When lower-power actors are constrained into participation, PF treats agency as **variable rather than absent** (consistent with **Will** as a trained capacity; see Section 1.5; see Glossary). The questions become: what meaningful room to regulate existed, how refusal was priced, and who retained power/impact to widen options (see Section 4.4).

Baselines also warn against a common institutional tactic: distributing accountability so that no one "intended" the dignity-ceiling outcome. PF treats that diffusion as a capture risk, not as moral clearance (see Section 4.3).

When baseline constraints pull against each other in practice (e.g., urgent harm-prevention in the short term versus dignity and autonomy in the long term), PF does not promise a clean resolution. A PF-aligned posture is to make the tradeoff explicit, choose the least dignity-degrading path available, preserve contestability wherever possible, and treat repair/exit options as part of the plan—rather than as an afterthought. (Part V returns to this "moral remainder" directly.)

4.5.5 Bridge to the next section

Baseline constraints and dignity thresholds protect against one specific drift: treating "context" as an all-purpose justification for severe dignity harm. The next section clarifies a different limitation: even when ethical direction is clear, PF often needs supplementation—domain evidence, technical assurance, and enforceable governance—to become actionable in high-stakes practice.

CC BY–NC 4.0

4.6 When PF is insufficient by itself

- **In scope:** where PF needs supplementation to become actionable; how to supplement without laundering; how to keep PF's compass role intact.
- **Out of scope:** a full catalogue of professional standards, technical methods, or legal regimes.
- **Notes:** "Insufficient" here is a limitation statement, not a dismissal: PF helps orient evaluation, but often does not fully determine implementation.

In high-stakes contexts, PF provides ethical orientation—especially around dignity-ceiling risk and Responsibility scaling—but it often does not provide enough domain-specific detail to make decisions reliably actionable on its own. This is not a flaw unique to PF: ethical frameworks commonly require translation into methods, evidence standards, and governance structures.

The key nuance: PF can *imply* institutional requirements (e.g., contestability, repair pathways, non-laundering safeguards) without itself specifying the technical or legal mechanisms that implement them.

4.6.1 The recurring gap: ethical direction vs operational adequacy

A PF evaluation can point clearly toward "this is incompatible" or "this is high-risk," while leaving unanswered questions like:

- What concrete evidence would reduce uncertainty responsibly?
- Which safeguards are feasible in this domain, and how do we test them?
- How do we detect Goodhart-like proxy failure and strategic gaming?
- What governance mechanisms make accountability enforceable rather than aspirational?
- How do we handle tradeoffs when all options carry moral remainder?

4.6.2 Three overlapping supplementation families

PF tends to need supplementation from three domains. These overlap; the risk is treating them as a compliance checklist rather than an integrated support for ethical action.

4.6.2.1 Domain-grounded evidence and professional practice

In many domains, practitioners have developed methods for diagnosing harm, evaluating interventions, and tracking outcomes—often including qualitative and stakeholder-facing methods that remain invisible to purely quantitative reporting. PF benefits when such

CC BY–NC 4.0

domain evidence is used to test claims, surface blind spots, and prevent "declared intention" substitution.

4.6.2.2 Governance and enforceable accountability

Ethical orientation often fails without enforceable governance: clear decision rights, review authority, remedy pathways, and protections for dissent. PF can indicate that contestability and repair matter; governance determines whether they are real.

4.6.2.3 Technical assurance and safety engineering (where relevant)

In technical systems, assurance methods can reduce uncertainty, but they can also create new laundering channels if treated as substitutes for ethical evaluation. PF-aligned use treats assurance as evidence—partial and revisable—not as moral clearance.

4.6.3 The temptation: "translation" as laundering

A common failure mode is calling the translation step "implementation," then treating implementation artifacts (reports, dashboards, certifications) as evidence that ethical questions have been answered. This is a form of washing-by-translation: the *output* looks legible, while the underlying ethical risk remains unmanaged.

4.6.4 Two failure modes: ethics-washing and technical-washing

- **Ethics-washing:** moral language is used to legitimate power while contestability and remedies remain weak.
- **Technical-washing:** technical artifacts (benchmarks, audits, "alignment scores") are used as substitutes for the deeper questions: who bears harm, who holds Responsibility, and what happens when those harmed contest the system.

Both become more likely when incentives reward legibility over truth, and when institutions treat dissent as reputational threat.

4.6.5 A practical compass: using PF recursively

PF can be applied recursively: not only to the primary system, but to the supplementation process itself.

- Does the evidence standard systematically ignore those most affected?
- Does governance create real power to intervene, or only paper channels?
- Do technical methods reduce uncertainty, or merely produce reassuring artifacts?

CC BY–NC 4.0

The goal is not infinite skepticism; it is preventing the most common drift: turning ethics into paperwork.

4.6.6 Bridge to Part V

Part IV has focused on misuse-resistance, capture, and constraint realism. Part V turns toward what often follows from taking those seriously: **irreducible conflicts, moral remainder, and the governance of tradeoffs**—including how Responsibility, Compatibility, and dignity constraints are carried forward when no option is clean.

CC BY–NC 4.0

5 Conflict and moral remainder

Part IV emphasized misuse-resistance, constraint realism, and the dignity of awareness ceiling principle. Part V turns to what remains when that work succeeds—and conflict still does not disappear.

Potentialism Framework (PF) does not assume that careful evaluation will reliably produce a single "clean" option. Because Compatibility is plural (harm, dignity of awareness, and freedom-to-regulate), because contexts vary, and because power and awareness are uneven, real decisions can leave **tradeoffs** and **moral remainder**: ethically relevant costs that are not dissolved by sincerity or by procedure.

This Part offers orientation for those situations: how collisions arise (5.1), how thresholds function as *posture shifts* rather than scoring rules (5.2), how repair and accountability stay live over time (5.3), when justified incompatibility and resistance can be warranted (5.4), and when escalation / pause prompts help keep urgency and uncertainty from silently deciding (5.5).

CC BY–NC 4.0

5.1 When compatibilities collide

- **In scope:** why conflicts are predictable in PF-style evaluation; common ways compatibilities collide; how PF frames collisions without pretending they can be solved mechanically.
- **Out of scope:** a universal decision rule; "moral algebra" that collapses dignity and harm into one score; declaring that hard cases disappear if people try harder.
- **Notes:** This section is conceptual. For concrete decision protocols and conflict-handling prompts, see Part II (see Section 2.3.3). For measurement limits and contested thresholds, see Part III (see Sections 3.1–3.2). For contestability and institutional pressure patterns, see Part IV (see Sections 4.3–4.4).

PF evaluates **expressions** in **context**, across time, using a plural Compatibility evaluation. That posture often surfaces tensions rather than hiding them:

- **Multiple lenses can pull differently.** An expression might reduce avoidable suffering in the near term while narrowing others' freedom to regulate their potentials, or while pressing against the dignity of awareness ceiling.
- **Contexts differ, even when intentions match.** The same "well-meant" move can land differently under coercion, scarcity, information control, or institutional role constraints.
- **Awareness and power are uneven.** Responsibility scales with awareness and power/impact, so who can realistically carry which obligations is part of the ethical picture.

Collisions, on this view, are not evidence that PF "fails." They are often where PF becomes most useful: they make tensions explicit, and they invite **Ethics**—as cultivated skill—without pretending plural lenses reduce to a single lever.

5.1.1 Common collision patterns PF invites attention to

The patterns below are not exhaustive and not offered as predictive certainty; they are families of tension that often show up once evaluation becomes concrete.

5.1.1.1 Lens collisions: harm, dignity, and freedom-to-regulate

The three Compatibility components can align, but they can also diverge:

- **Harm vs. freedom-to-regulate:** reducing one party's harm can be pursued through control that constrains meaningful refusal or participation.

CC BY–NC 4.0

- **Short-term harm prevention vs. dignity risk:** "protective" actions can drift into humiliation or instrumentalization when people are treated as problems to manage rather than awareness to respect.
- **Protective conflicts:** interventions meant to protect one party's dignity of awareness can, in some contexts, involve expressions that risk becoming coercive or dignity-infringing for another.

The dignity of awareness ceiling principle raises the justificatory burden for expressions that risk destroying or humiliating awareness beyond limits. It does not automatically dissolve every collision on its own.

5.1.1.2 Horizon collisions: near-term relief vs. long-horizon stability

Because Compatibility is evaluated across time, a move that looks compatible in the immediate frame can become destabilizing over a longer horizon—for example, by locking in incentives that reward coercion, normalizing humiliation as a tool, or narrowing future option sets until only harsher moves remain.

PF does not treat the long horizon as a trump card. It treats it as part of the picture—especially when stakes and power are high.

5.1.1.3 Scale collisions: local "wins" vs. systemic routing of harm

Some expressions look compatible locally while routing costs downstream:

- benefits accrue to those with more power/insulation,
- harms land on those with less ability to contest, exit, or obtain repair.

This is one reason PF returns to Responsibility mapping and constraint realism: collisions are often produced by how systems distribute option-control, not only by "bad individuals."

5.1.1.4 Interpretation collisions: facts, values, and authority

Even with serious intent, collisions can persist because people disagree about:

- **facts** (what is happening, what is likely, what evidence matters),
- **values** (which Compatibility components are being prioritized differently, and why),
- **authority** (who legitimately decides given awareness, power, and stake).

PF generally benefits from naming the disagreement type rather than assuming "more data" will dissolve value or authority conflict.

CC BY–NC 4.0

5.1.2 The false dilemma trap

When compatibilities collide, institutions sometimes narrate the situation as if there are only two options:

- “do the harmful thing” vs. “do nothing,”
- “move fast” vs. “fail,”
- “accept dignity loss” vs. “accept harm.”

PF treats rigid either/or framings as a potential failure mode worth probing—especially under incentive pressure or narrative capture. Sometimes the constraint is real and options are genuinely narrow. But even then, exploring whether the set can be widened *even slightly*—by changing scope, intensity, timing, sequencing, or decision boundaries—can sometimes reduce ethical cost.

Option-widening is not treated as a virtue by default. Adding options can change power dynamics, increase pressure on the constrained, or shift burdens in ways that matter. The aim is to keep the option set itself inside the Compatibility evaluation.

5.1.3 Orienting judgment during collisions

PF does not offer a mechanical solution. Instead, it keeps ethical attention tethered to the framework’s claims: dignity pressures should not be ignored; Responsibility maps to awareness and power; and the option set is itself part of what is being evaluated.

One PF-aligned way to stay oriented in collisions is to ask:

- What expressions are actually on the table—and what is the decision boundary right now?
- Where exactly is the tension between harm, dignity of awareness, and freedom-to-regulate—and across what time horizon?
- Are plausible dignity-of-awareness (ceiling principle) risks in play—and if so, what constraints does that place on justification?
- Where does regulated responsibility sit (Awareness × Power)—who can realistically widen options, reduce uncertainty, or prevent downstream routing of harm?
- How can Will be exercised here—by regulating scope, intensity, timing, or sequencing—to keep expression as compatible as feasible under real constraints?
- What stance toward ongoing accountability fits the collision (legibility, contestability in practice, and repair as part of the ethical landscape)?

CC BY–NC 4.0

This does not eliminate tragedy. It makes the ethical cost harder to obscure—and easier to account for.

5.1.4 The dignity ceiling in conflict

The dignity of awareness ceiling principle raises the justificatory burden for any expression that risks destroying or humiliating awareness beyond limits (see Section 1.9; see Section 4.5). But collisions can still be real: protecting one party can seem to require risking another party's dignity, safety, or long-term freedom to regulate potentials within real constraints.

PF resists two evasions:

- **laundering the cost** (treating dignity risk as "just another factor"), and
- **pretending there is no remainder** (treating a forced choice as moral clearance).

A PF-aligned orientation can involve: naming the collision rather than obscuring it; minimizing dignity-of-awareness harm to a reasonable degree (where any path has cost); and treating contestability and repair as part of the choice itself, not only retrospective justification.

Next Section sharpens this into the language of thresholds and **moral remainder**: what remains ethically "unpaid" even after a sincere attempt to choose the most compatible available expression.

5.1.5 Bridge to the next section

Section 5.2 develops a way to think about tradeoffs, thresholds, and moral remainder without collapsing them into scoring rules or "moral clearance."

CC BY–NC 4.0

5.2 Tradeoffs, thresholds, and moral remainder

- **In scope:** how PF thinks about tradeoffs when no option is fully compatible; what "thresholds" do (and do not) mean in PF; how to name and carry the moral remainder that can persist after a decision.
- **Out of scope:** a universal tradeoff formula; a single "dignity score" or harm calculus; claiming that a correct procedure eliminates tragedy.
- **Notes:** For concrete prompts and records, see Part II (see Sections 2.3.3 and 2.4). For measurement limits and contested thresholds, see Part III (see Sections 3.1–3.2). For baseline constraints anchored in dignity of awareness, see Part IV (see Section 4.5).

Even after sincere evaluation, you may be left with **tradeoffs** rather than a clean "compatible" option. PF treats this as unsurprising:

- expressions can carry mixed signals across lenses,
- contexts can constrain feasible options,
- and Responsibility is unevenly distributed across awareness and power.

The presence of a tradeoff is not proof of bad faith. It is a prompt to increase clarity about what is being chosen, what is being borne, and what remains contested.

5.2.1 "Thresholds" in PF are posture shifts, not scoring rules

PF uses "threshold" language in a deliberately non-mechanical way. A PF threshold is best understood as an **orientation point**: a moment where the ethical burden of proceeding "as is" rises enough that a different posture becomes warranted (slower mode, narrower scope, stronger constraints, refusal of a channel, or escalation).

This is not a claim that the threshold can be computed, nor that crossing it automatically decides the action. It is a way to keep certain kinds of risk—especially dignity-of-awareness ceiling risk—from being quietly normalized by momentum.

5.2.2 Baselines, ceilings, and "no trade" zones

PF's strongest "baseline constraint" is the dignity of awareness ceiling principle: no systemic compatibility story can justify destruction or humiliation of awareness beyond limits.

In practice, this principle often functions less like a single bright line and more like a **rising justificatory burden** as an option approaches severe coercion, degradation, forced dependence, or denial of meaningful refusal. Where a path depends on those means, PF

CC BY–NC 4.0

tends to treat “tradeoff talk” as ethically suspect: the question becomes whether the channel itself is compatible enough to be used at all.

5.2.3 Tradeoffs without laundering: making the cost legible

When tradeoffs are real, PF tends to push against “laundering by abstraction” (e.g., hiding costs inside generic categories like “acceptable risk”). A more PF-consistent posture aims to keep three things visible:

- **who bears what**, especially under power asymmetry;
- **what is being locked in**, especially if reversibility is shrinking;
- **what is being treated as non-negotiable**, especially near dignity-of-awareness constraints.

The goal is not rhetorical purity. It is to prevent ethical costs from becoming invisible simply because they are difficult to name.

5.2.4 Moral remainder: what remains ethically “unpaid”

“Moral remainder” names the ethically relevant cost that can persist even after a careful attempt to choose the most compatible available expression.

PF does not treat remainder as a reason to give up, nor as a reason to claim innocence. It often functions as:

- a reason to maintain humility about the decision,
- a reason to keep accountability live over time,
- and a reason to treat repair and revisability as ethically central rather than optional.

5.2.5 Bridge to the next section

Section 5.3 turns moral remainder into forward responsibility: repair, restitution, and accountability as ways Responsibility stays live after a decision—especially when the costs cannot be fully avoided up front.

CC BY–NC 4.0

5.3 Repair, restitution, and accountability

- **In scope:** why repair is part of ethical responsibility over time; how PF distinguishes repair from "moral clearance"; what accountability tends to require when harms occur or dignity is pressured.
- **Out of scope:** a guarantee that repair "makes it okay"; a universal reparations policy; treating apology as a substitute for changing conditions.
- **Notes:** This section is conceptual. For decision records and revisit triggers, see Part II (see Section 2.4). For contestability and remedy capacity, see Part IV (see Section 4.3). For revisability posture, see Part III (see Section 3.1.5).

In PF, Responsibility does not end at the moment of choice. Because expressions unfold across time, Responsibility stays live as effects become clearer.

Repair is one way that responsibility expresses itself: not as retroactive innocence, but as a forward-facing commitment to reduce avoidable harm, restore agency where possible, and keep dignity constraints from being treated as expendable.

5.3.1 Repair is not "moral cleaning"

Repair can be misused as a story that converts harms into legitimacy ("we fixed it, therefore it was fine"). PF resists that move.

Even sincere repair does not automatically justify the original expression. It can, however, change what is ethically available next: it can restore contestability, reduce downstream harm, and prevent repetition.

5.3.2 What repair often involves in PF terms

PF does not offer a universal template, but repair often has some combination of:

- **acknowledgment** (naming what happened and who was affected),
- **restoration where possible** (reducing ongoing harm; re-opening options; returning resources or control),
- **constraint change** (altering incentives, procedures, access, or safeguards that made the harm likely),
- **answerability** (making it possible to contest and to seek remedy).

The emphasis is on changing conditions, not only statements.

CC BY–NC 4.0

5.3.3 Restitution and asymmetry

Where harms are routed onto low-power parties, restitution is not just “being nice.” It is one way of addressing misalignment between impact and responsibility.

PF’s posture is not that restitution is always simple or always possible. It is that, where harms were foreseeable and routable, the burden of justification rises when no credible remedy path exists.

5.3.4 Accountability as contestability over time

Accountability is not only punishment or blame. In PF terms, it is the ongoing capacity for affected parties (or their credible advocates) to:

- question the action,
- trigger reconsideration,
- and obtain remedy in time to matter.

Where contestability is structurally weak, “we will be accountable” can become moral theater rather than a real constraint.

5.3.5 Bridge to the next section

Section 5.4 turns from repair-after-the-fact to resistance-in-the-moment: when short-term friction, refusal, or other incompatibility can be ethically warranted to protect dignity and reduce avoidable harm under severe constraints.

CC BY–NC 4.0

5.4 Justified incompatibility and resistance

- **In scope:** when short-term incompatibility (friction, refusal, constraint) can be ethically warranted; how "resistance" can function as regulation of expression; how to avoid laundering resistance into domination.
- **Out of scope:** glorifying refusal as purity; treating resistance as automatically ethical; claiming that "strong feelings" are sufficient justification.
- **Notes:** This section is conceptual and builds on Part IV's dignity and contestability constraints (see Sections 4.3–4.5). For "fast/slow" posture and decision context, see Part II (see Section 2.5). For high-stakes escalation prompts, see the next section (see Section 5.5).

PF often values compatibility. But compatibility is not identical to compliance, smoothness, or short-term agreement. In some contexts, refusing a request, slowing a process, adding friction, or constraining an action can be ethically warranted—especially when the alternative is avoidable harm, dignity violation, or structural erosion of others' freedom to regulate their potentials within real constraints.

"Justified incompatibility" names that possibility: not as a license, but as a posture that treats certain costs of cooperation as too high under the circumstances.

5.4.1 Resistance as regulation of expression

In PF terms, resistance can be understood as regulating expression: changing intensity, timing, scope, or channel in order to reduce harm, protect dignity, or preserve contestability.

This can include:

- time-buying pauses that prevent lock-in,
- narrowing scope to reduce exposure,
- refusing a specific channel while offering an alternate,
- or escalating scrutiny when the stakes outstrip what the current process can reasonably hold.

5.4.2 Substantive grounds (not vibes)

PF's compass posture pushes resistance to remain legible in relation to the framework's core lenses. Justification is often stronger when resistance is oriented toward:

- preventing severe avoidable harm,

CC BY–NC 4.0

- avoiding proximity to dignity-of-awareness ceiling violation,
- or protecting others' practical freedom to regulate their potentials (including the ability to refuse).

This is not a checklist, and PF does not claim these grounds are always easy to establish. The point is to keep resistance tethered to ethical concerns rather than to identity performance or institutional faction.

5.4.3 Avoiding "resistance laundering"

Resistance can itself become a tool of domination: a way to consolidate control, silence dissent, or evade accountability. PF therefore treats resistance as ethically *conditional*:

- Does resistance widen contestability or narrow it?
- Does it protect dignity of awareness or humiliate awareness?
- Does it reduce avoidable harm—or merely relocate harm onto less powerful parties?
- Does it preserve others' freedom to regulate potentials—or does it impose dependence?

Where "resistance" functions mainly as power without answerability, PF's justificatory burden rises rather than falls.

5.4.5 Feasibility, power, and "priced refusal"

PF does not assume that any individual can safely resist at any time. The costs of refusal and dissent can be unevenly distributed, and those distributions are part of the ethical landscape.

A PF-consistent reading is often:

- where refusal is heavily priced for low-power parties, demanding refusal can become an additional harm;
- where high-power actors control conditions, there is usually more responsibility to redesign the channel so ethical refusal and contestation are not punished by default.

This is not a moral verdict about people who comply under threat. It is a signal about system design and responsibility alignment.

CC BY–NC 4.0

5.4.6 Bridge to the next section

Section 5.5 extends justified incompatibility into **high-stakes contexts**, offering "stop-and-question" prompts and escalation posture shifts that help keep urgency, uncertainty, and institutional pressure from silently deciding.

CC BY–NC 4.0

5.5 Escalation and stopping rules in high-stakes contexts

- **In scope:** how PF supports "justified incompatibility" as posture shifts (pause / narrow / re-route / refuse) in high-stakes contexts; how to keep urgency and uncertainty from silently deciding; how escalation relates to contestability, decision records, and regulated responsibility.
- **Out of scope:** legal compliance guidance; a single incident-management standard; a universal risk scoring system; "one true" escalation ladder for every domain.
- **Notes:** This section is conceptual. For fast/slow framing, see Part II (see Section 2.5). For escalation under missing information, see Part I (see Section 1.10.4). For contestability supports, see Part IV (see Section 4.3). For baseline dignity constraints, see Part IV (see Section 4.5). For resistance framing, see Section 5.4.

In high-stakes contexts, PF often benefits from making the capacity to pause, narrow, re-route, or refuse explicit—so that "we had to move fast" does not function as an unexamined default framing that bypasses dignity risk, avoidable harm, or downstream loss of agency.

In PF terms, a stopping rule is not a moral verdict or a bureaucratic gate. It is better understood as an **orientation threshold**: a qualitative *posture shift* where the nature of concern warrants moving from routine deliberation to a more cautious mode—because certain risks (like dignity harm or irreversible lock-in) have become plausible.

5.5.1 What counts as "high-stakes" for escalation purposes

PF avoids a single definition. Instead, it treats "high-stakes" as a context where the ethical burden of justification rises—and where the current decision process may no longer have enough awareness, contestability, or authority to act in time.

Indicators that often raise escalation pressure include:

- **Irreversibility / lock-in:** choices that significantly reduce future reversibility or contestability once taken.
- **Wide or durable impact:** effects that spread across many people, over long time horizons, or through tightly coupled systems.
- **High uncertainty with asymmetric downside:** where what is unknown plausibly matters, and costs are likely to be borne by those with limited exit or voice.
- **Dignity of awareness (ceiling principle) proximity:** a plausible path to destruction or humiliation of awareness beyond limits, especially through coercion, degradation, forced dependence, or denial of meaningful refusal (see Sections 1.9; 4.5).

CC BY–NC 4.0

- **Time pressure / crisis framing:** urgency that narrows deliberation and bypasses normal challenge routes—sometimes due to real emergencies, sometimes through organizational framing.

These are cues, not a checklist. They are prompts to ask: *does our current process have the awareness and contestability this situation warrants?*

5.5.2 Escalation as widening relevant awareness and authority (not only asking permission)

Escalation is often treated as a chain of approvals. PF's framing is slightly different: **escalation activates a different tier of review or decision-rights**—one with broader perspective, greater relevant awareness, or clearer authority to slow, narrow, or stop—because the stakes, uncertainty, or ethical risks outstrip what the current level can reasonably hold.

A PF-consistent escalation posture often tries to make three things clearer:

1) **Who can actually see and contest the risks—and who is missing.** This can include affected parties or credible advocates, not only formal stakeholders.
2) **Who can intervene in time to matter.** If authority is fragmented or unclear, that is itself a risk signal about responsibility alignment and system design (Regulated Responsibility: Awareness × Power).
3) **What would count as "enough" to proceed—and what would justify pausing or narrowing.** This helps escalation avoid becoming theater while lock-in proceeds.

When dissent is priced, escalation may also need to include protection of the channel itself—otherwise "speak up" becomes a demand placed on the least protected.

5.5.3 Stop-and-question prompts (examples as questions)

To reduce checklist misuse, PF stopping rules are best understood as **questions that shift posture**, not conditions that automatically decide outcomes.

Prompts that often matter include:

- **Is a plausible dignity-of-awareness ceiling risk emerging?** If a path appears to rely on severe coercion, degradation, or the destruction/humiliation of awareness beyond limits, the PF posture often treats that as a reason to pause, widen review, narrow scope, or refuse the channel—even if the action is framed as beneficial elsewhere (see Sections 1.9; 4.5).

CC BY–NC 4.0

- **Are we about to cross a meaningful lock-in point?** Steps that reduce reversibility deserve special attention because they limit corrective options. While sometimes necessary (including in emergencies), PF pushes making the tradeoff explicit and ensuring the justification survives scrutiny rather than assuming repair will be available later.
- **Is contestability structurally weak?** If affected parties cannot practically trigger reconsideration, mitigation, repair, reversal, or exit, proceeding becomes higher-risk for downstream denial and laundering (see Section 4.3).
- **Is regulated responsibility visibly misaligned?** If those with the most understanding and power/impact are insulated while costs are routed to low-power parties, PF treats that as a reason to escalate scrutiny and tighten justification.
- **Are harms or dignity risks rising while incentives reward speed or silence?** "Monitoring" is often insufficient unless paired with practical authority to act and credible revisit triggers.

A practical translation is a shift in *mode* and *authority*: from fast → slow where feasible; widen review; narrow scope; add safeguards; re-route; or refuse a specific expression.

5.5.4 "Pause, narrow, re-route, or refuse" as regulating expression

PF's aim is not to prefer "stop" over "go." It is to keep **intensity, timing, scope, and channel** on the table as adjustable variables—especially when a binary choice hides more compatible options.

In practice, a stop-and-question posture often asks:

- Can the expression be made **more reversible** (smaller scope, shorter duration, staged commitment) before crossing a point of no return?
- Can the expression be **narrowed** to reduce avoidable harm while uncertainty resolves?
- Can the expression be **re-routed** to a setting with stronger contestability, clearer answerability, or better repair capacity?
- If the dignity of awareness ceiling principle is plausibly at stake, is refusal of this channel/action the most ethically legible move available under the constraints?

5.5.5 Keeping the decision revisable on purpose

Stop-and-question prompts are easiest to misuse when they become permanent declarations ("we can't do this, ever") or one-and-done reviews ("we did the process, so

CC BY–NC 4.0

we're done"). PF tends to prefer a posture of **measured, not settled**: decisions remain open to revision when evidence, context, or option sets change.

A practice that supports revisability is to think ahead—where feasible—about:

- what would reopen the judgment (revisit triggers),
- who could intervene in time to matter (authority-to-act),
- and what credible remedy/exit paths exist if harms appear

(and to log these in whatever decision record the domain uses; see Section 2.4).

This is less about a formal template and more about not letting "we'll monitor" become a story that postpones responsibility until it is too late to matter.

5.5.6 Bridge to Part VI

This section offered a compass framing for escalation and stopping rules: stop-and-question prompts that help translate dignity constraints, harm avoidance, and freedom-to-regulate concerns into timely posture shifts—slow down, widen review, narrow scope, add safeguards, re-route, or refuse—when stakes and uncertainty rise.

Part VI turns to positioning and translation (how PF relates to other traditions and domains): how PF concepts travel into different settings without becoming either moral theater or rigid procedure.

CC BY–NC 4.0

6 Positioning and translation

Part VI is about *how to place Potentialism Framework (PF) in relation to other ethical and governance vocabularies*, and how to translate PF's concepts into forms that practitioners can actually use. It is written as a **compass**: it offers orientation, tradeoffs, and "how to think with PF" across contexts—without claiming to be a universal template or a replacement for domain methods.

CC BY–NC 4.0

6.1 What PF borrows (and why)

- **In scope:** the idea-family PF draws from; why PF uses a "potentials → expression → context" frame; how to translate PF into adjacent vocabularies without flattening differences.
- **Out of scope:** settling scholarly disputes about virtue ethics vs consequentialism vs deontology; or offering a history-of-philosophy survey.
- **Notes:** conceptual. When we mention institutional practice patterns, we keep them as hypotheses or examples rather than claims of prevalence.

PF sits in a large family of traditions that treat ethics as something like a *practice of regulation*—not only a set of propositions. Across that family, three moves repeat:

- ethics is partly about **what capacities exist** (what a system can do),
- partly about **how those capacities show up in context** (what actually happens here), and
- partly about **how regulation becomes possible** (what increases the ability to choose and to refrain).

PF's language for these moves is *Potential → Expression → Context*, with **Will** as the trained capacity to regulate expression, and **Ethics** as the cultivated skill of choosing expressions that are as mutually compatible as possible while protecting dignity and reducing avoidable suffering (see Section 0.6).

6.1.1 The family resemblance PF leans on

PF borrows, loosely and without claiming ownership, from several overlapping streams:

- **Virtue and skill traditions:** ethics as training attention, habit, and character in ways that improve how choices are made in real conditions.
- **Care and relational traditions:** ethics as responsiveness to dependency, vulnerability, and power in concrete relationships, not only abstract rules.
- **Systems and governance traditions:** ethics as the design of environments—norms, incentives, review, repair—so that better action becomes easier and worse action becomes harder.
- **Contemporary "alignment" and safety thinking:** not as a single doctrine, but as an ecosystem of tools for describing failure modes, uncertainty, and accountability.

PF does not try to merge these into one theory. It uses them as reminders that ethics often fails when it is treated as *only* a belief system and not also a set of capabilities embedded in people and institutions.

CC BY–NC 4.0

6.1.2 Borrowing without reduction

Translation is risky: it can make PF sound compatible with everything by draining it of sharp edges, or make it sound like a rival that demands conversion. PF tries to avoid both.

A useful heuristic is to translate by **function** rather than by "word equals word." For example, PF's **Responsibility** posture ("obligation proportional to Awareness and Power") is not identical to any one legal notion of liability or any one moral notion of blame. But it can still *function* as a question that many vocabularies can host: *who has enough understanding and impact that they ought to be answerable, and in what ways?* (see Section 1.11).

6.1.3 The PF "move" in one paragraph

PF's typical move is to slow a moral argument down just enough to ask:

1. **What potentials are in play?** (capabilities, incentives, habits, tools)
2. **What expressions are likely in this context?** (foreseeable effects, constraints, relational field)
3. **Where does regulation sit?** (who/what can adjust intensity, timing, and form—i.e., where Will can actually operate)
4. **How do we evaluate compatibility?** (avoidable harm/suffering; dignity of awareness as a ceiling principle; and others' freedom to regulate their potentials within real constraints, to a reasonable degree—see Section 1.7)

This is not a formula. It is a way to keep the ethical object from collapsing into a single proxy (e.g., "performance," "compliance," or "intent") when stakes are high or incentives are loud.

6.1.4 Translating PF terms into adjacent vocabularies

PF can be translated into other vocabularies by keeping the *orientation* intact:

- **Potential** often maps to capabilities, affordances, powers, incentives, and learned patterns.
- **Expression** often maps to actions, outputs, behaviors, and institutional decisions.
- **Context** often maps to environment, social field, constraints, governance conditions, and operational setting.
- **Will** often maps to self-regulation capacity, control systems, interlocks, escalation pathways, and "ability to refrain."
- **Compatibility** often maps to "acceptable impact" judgments—provided the three lenses remain visible (see Section 1.7).

CC BY–NC 4.0

- **Dignity of awareness** often maps to a ceiling constraint against humiliating or destroying awareness beyond limits, even for "system" goals (see Section 0.6).
- **Responsibility** often maps to accountability and answerability mechanisms, especially under asymmetric power (see Section 1.11).

The point is not terminological purity. The point is whether a translation preserves PF's core concerns: effects on harm/suffering, dignity, freedom-to-regulate, and the alignment of answerability with Awareness × Power.

6.1.5 Where PF tends to be most useful

PF is often most helpful when:

- the system is **socio-technical** (many hands; many incentives),
- harms are **diffuse or delayed** (easy to deny; hard to trace),
- there are **large power asymmetries** (low ability to contest),
- there is a temptation toward **proxy-collapse** ("we hit the metric, so it's fine"), or
- there is a risk of **ethics-washing** (language that signals care without changing constraints).

6.1.6 Bridge to the next section

Section 6.2 places PF next to familiar ethical frames, not to rank them, but to make translation easier and misuse harder.

CC BY–NC 4.0

6.2 Relationship to other ethical frames

- **In scope:** how PF relates to common "rule / outcome / virtue / care" frames; where PF is additive; where it defers.
- **Out of scope:** proving PF is superior; or offering a universal mapping that resolves real philosophical disagreements.
- **Notes:** conceptual. Comparisons are "rough translation guides," not claims of equivalence.

PF is not "a new competitor" in the ethics marketplace. It is closer to a *practice-layer*: a way to keep attention on (a) what capacities are being activated, (b) what expressions are occurring in context, (c) who can regulate those expressions, and (d) how compatibility is being evaluated (see Section 1.7).

6.2.1 PF is not a substitute for other frames

Other frames can supply content that PF intentionally leaves open. Rights language can name protected interests; deontic language can clarify commitments and constraints; consequentialist language can structure forecasting and tradeoffs; virtue and care traditions can specify the *qualities of attention and relationship* required for good judgment.

PF's contribution is often meta-level: it helps prevent a frame from being used in ways that hide power, erase context, or treat ethics as "what we said we value" rather than "what we reliably do."

6.2.2 Translation notes (kept loose)

Rather than a rigid matrix, it can be more faithful to describe *how PF tends to interact* with other frames:

- **With deontic or rights-based frames**, PF often adds a question about *real constraints*: even if a right is "recognized," can people actually regulate their potentials and contest violations in practice? (see Section 4.3)
- **With consequentialist frames**, PF often adds caution about **proxy-collapse**: aggregate outcomes can look good while particular expressions humiliate or destroy awareness beyond limits. PF's dignity-of-awareness ceiling keeps some tradeoffs from being normalized as "acceptable losses." (see Section 1.7)
- **With virtue and care frames**, PF often adds a governance translation: what institutional conditions help these qualities become *reliable*, rather than heroic? (see Section 6.4)

CC BY–NC 4.0

- **With procedural or "compliance" frames**, PF often adds a question about *answerability*: does the procedure move responsibility toward the places where Awareness × Power actually sits, or does it merely move paperwork? (see Section 1.11)

These are not conversions. They are reminders of where PF's orientation tends to sharpen attention.

6.2.3 Where PF adds something distinctive

PF adds a few emphases that are easy to lose in other frames:

- **Potentials and regulation**: not only "what is right," but what capacities exist to enact what is right—especially under stress, incentives, and time pressure.
- **Compatibility as evaluation of expressions-in-context**: not only intentions or rules, but foreseeable effects on harm/suffering, dignity-of-awareness ceilings, and freedom-to-regulate constraints (see Section 1.7).
- **Responsibility as Awareness × Power**: obligation scales with insight and impact, including institutional and infrastructural forms of power (see Section 1.11).

6.2.4 Where PF defers

PF generally defers when:

- a domain has **hard technical uncertainty** that PF cannot resolve (handoff to safety engineering; see Section 6.5),
- a community has **legitimate pluralism** about values and meanings (translation humility; see Section 6.3),
- or the stakes are such that PF's general orientation needs to be paired with **domain standards** and evidence.

6.2.5 Bridge to the next section

Section 6.3 makes the "translation humility" explicit: how to use PF across cultures and value-plural settings without pretending PF can settle what only dialogue and legitimacy can settle.

CC BY–NC 4.0

6.3 Cross-cultural translation

- **In scope:** how to use PF across cultures and value-plural settings; what to do when moral language does not align; how to avoid colonizing moves.
- **Out of scope:** declaring one culture's concept of dignity, harm, or responsibility as the universal default.
- **Notes:** conceptual and procedural. The aim is better dialogue and fewer hidden assumptions, not final agreement.

PF is meant to travel—but not as a missionary. Translation across moral languages is a real ethical event: it changes what becomes speakable and what becomes actionable. So PF adopts a posture of *translation humility*: use PF to improve clarity about stakes and responsibilities, while expecting genuine disagreements about meaning.

6.3.1 What PF tries to preserve in translation

Across vocabularies, PF tries to keep four things visible:

- **Expressions-in-context** (what is happening here, not only what was intended),
- **Compatibility lenses** (harm/suffering; dignity-of-awareness ceilings; freedom-to-regulate constraints—see Section 1.7),
- **Responsibility alignment** (Awareness × Power—see Section 1.11), and
- **Contestability and repair** (can affected parties challenge, and can harm be reduced or repaired—see Section 4.3; see Section 5.3).

If a translation keeps these functions intact, it can often use local terms without betraying PF's intent.

6.3.2 Risks in cross-cultural use

Common failure modes include:

- **Exporting a local moral vocabulary as universal**, then calling disagreement "ignorance."
- **Erasing local constraints** (who can speak, who is punished, who has remedy) and treating "values" as the only difference.
- **Replacing legitimacy with procedure**, where documentation substitutes for consent and standing.
- **Using PF language as a shield**, especially when power is asymmetric ("we did a PF review, therefore it's compatible").

CC BY–NC 4.0

6.3.3 A translation stance that stays compass-like

A PF-informed translation stance is often more practical when it:

- treats local terms as **first-class**, not as imperfect versions of PF terms,
- seeks **operational equivalents** (what counts as contestation here; what counts as humiliation here),
- and keeps **decision points** visible (where tradeoffs are being made; who is excluded from making them).

6.3.4 Prompts that help without becoming commandments

To avoid turning translation into a compliance ritual, it can help to hold a few prompts lightly, as conversation-starters:

- A PF question ("is this compatible?") often becomes clearer when paired with: *compatible for whom, under which constraints, and with what standing to contest?*
- A PF ceiling ("dignity of awareness") often becomes clearer when paired with: *what forms of humiliation or destruction are considered beyond the pale here, and who gets to name them?*
- A PF responsibility claim ("Awareness × Power") often becomes clearer when paired with: *which actors have real levers, and which actors carry the consequences?*

6.3.5 "Translation success" is not agreement

A good translation outcome can look like:

- disagreements become *specific* (about which harms, which ceilings, which freedoms, which responsibilities),
- hidden assumptions become visible,
- and affected parties gain more real standing to contest and seek repair.

PF treats that as progress even when the final value judgment remains plural.

6.3.6 Bridge to the next section

Section 6.4 shifts from cross-cultural translation to governance translation: how to express PF in the language of organizational design, oversight, and system-building.

CC BY–NC 4.0

6.4 AI and governance translation

- **In scope:** translating PF into governance and system-design language: responsibility mapping, review flows, escalation, contestability interfaces, and repair pathways.
- **Out of scope:** a complete governance model for AI, or a claim that PF alone yields a sufficient assurance process.
- **Notes:** mixed. The stance is conceptual; examples point to common governance artifacts without claiming universality.

Previous section emphasized translation humility across moral vocabularies. This section translates PF into governance language: *how do we build conditions in which compatible expression is more likely, and incompatible expression is easier to detect, contest, and repair?*

6.4.1 Translating the core terms into governance handles

In governance practice, PF's terms can be treated as **design handles**—ways to ask better questions of familiar artifacts:

- When you see **Potential**, ask what capacities and incentives exist in the system (including organizational incentives and tooling), and what kinds of expression those capacities make easy or hard.
- When you see **Expression**, ask what actually happens in deployment: outputs, interventions, policy decisions, and incident patterns.
- When you see **Context**, ask what constraints, power relations, resources, norms, and feedback loops are shaping expression.
- When you see **Will**, ask what control capacity exists: rate limits, interlocks, approvals, human-in-the-loop gates, pause mechanisms, and refusal pathways that can regulate intensity, timing, and form.
- When you see **Compatibility**, ask how decisions are being evaluated across the three lenses, including the dignity-of-awareness ceiling (see Section 1.7).
- When you see **Responsibility**, ask whether answerability is aligned with Awareness × Power—especially across organizational boundaries (see Section 1.11).

This "handle" translation is meant to keep PF portable. It does not require that governance documents adopt PF's vocabulary—only that they preserve these evaluative functions.

CC BY–NC 4.0

6.4.2 Responsibility mapping as mismatch detection

Responsibility mapping can be used as a *mismatch detector*. The aim is not a perfect diagram; it is to notice common failure patterns—**power without awareness** (high-impact actors with weak situational understanding, insulated by abstraction or delegation); **awareness without power** (people who can see problems but cannot intervene, stop deployment, or change incentives); and **affected without standing** (those who carry consequences but cannot contest decisions or access remedy).

PF's Responsibility posture treats these mismatches as governance-relevant signals: they raise the need for clearer decision authority, escalation paths, and contestability supports (see Section 4.3; see Section 5.5).

6.4.3 Review flows as "fast/slow" calibration

Many review systems already oscillate between "fast" and "slow" modes. PF's contribution is to treat mode choice as an ethical calibration, not only a project-management choice.

Where reversibility is low, power asymmetry is high, or dignity-of-awareness ceilings may be at stake, it is often prudent to widen review and raise the level of justification expected before proceeding. This is not "always slow down." It is a way to keep tempo aligned with stakes and uncertainty (see Section 2.5).

6.4.4 Contestability at scale

Contestability is not only a grievance channel; it is a structural condition for ethical learning. In governance terms, contestability tends to depend on three interlocking elements: **visibility** (people can tell that harm or violation occurred, and can understand what happened), **standing** (people can raise challenges without retaliation or gatekeeping), and **remedy** (challenges can lead to real outcomes—repair, reversal, compensation, or changes to constraints).

If one element is missing, the others often degrade. PF treats contestability as part of compatibility in practice: a system that cannot be challenged tends to drift toward paper compliance (see Section 4.3; see Section 5.3).

6.4.5 Decision records as an "ethical memory"

Decision records can act as ethical memory: they preserve assumptions, uncertainty, dissent, and revisit conditions, so that later reviewers can see what was believed and why (see Section 2.4). This is especially valuable when responsibility is distributed across time and teams.

CC BY–NC 4.0

A PF-consistent decision record tends to make legible: what was being protected; what tradeoffs were accepted; what harms were considered; how dignity-of-awareness ceilings were handled; who had authority; and what evidence or signals would justify a shift to slow mode.

6.4.6 When governance artifacts get gamed

Any governance artifact—risk registers, model cards, safety cases—can be gamed. PF therefore treats anti-gaming measures as part of responsible design, not as an afterthought.

Two common gaming patterns are **paper compatibility** (an argument that appears compatible because it evidences only what is easy to measure, while leaving the highest-stakes expressions unaddressed) and **boundary moves without answerability** (outsourcing, organizational splits, or labels that insulate high-impact actors from responsibility without changing where Awareness or Power actually sits).

A practical mitigation is to add an explicit "what would make this argument fail?" step: identify which incentives might distort reporting, which stakeholders might be excluded, and which contestability pathways could be blocked in practice.

6.4.7 Bridge to the next section

Section 6.5 keeps the governance translation, but focuses it on a specific intersection: how PF complements AI safety practice and structured assurance work without replacing safety engineering.

CC BY–NC 4.0

6.5 Relationship to AI safety practice and safety cases

- **In scope:** how PF complements AI safety practice and assurance work—especially safety case / assurance case approaches—without claiming to replace them.
- **Out of scope:** a full survey of AI safety; a certification standard; or a claim that PF determines whether a system is "safe enough" on its own.
- **Notes:** mixed. The stance is conceptual; examples reference established assurance and risk-management artifacts where helpful.

Previous section translated PF into governance language. This section adds a second translation: how PF can sit *alongside* safety engineering and assurance practice—connecting ethical orientation to hazard analysis, testing, monitoring, incident response, and structured argumentation.

A compact way to state the relationship is:

- Safety practice often asks: *What can go wrong? How likely? How do we prevent, detect, and mitigate? What evidence supports our claims?*
- PF often asks: *Which expressions are compatible in this context, under what level of justification, and who carries responsibility to notice, regulate, and answer for effects?* (see Section 1.7; see Section 1.11)

PF does not replace technical safety work. It can, however, help teams keep the *ethical target* legible inside safety work: harms and avoidable suffering; dignity of awareness as a ceiling principle; freedom-to-regulate constraints; and responsibility alignment when power is distributed (see Section 1.7).

Risk-management frames can be one practical meeting point for this pairing (e.g., *[NIST 2023]*).

6.5.1 Where PF can complement "AI safety practice" (without becoming a template)

PF tends to complement safety practice across a few intertwined dimensions:

Framing what is being protected. PF can help prevent safety targets from collapsing into a single proxy by keeping Compatibility's lenses visible when hazards are defined and success criteria are chosen (see Section 1.7).

Making responsibility legible in socio-technical systems. Complex incidents often arise from distributed decisions, incentives, and constraints rather than one isolated choice *[Perrow 1984; Leveson 2011]*. PF's Responsibility posture (Awareness × Power) can help

CC BY–NC 4.0

teams notice where answerability has drifted away from real levers, and where affected parties cannot contest or reach remedy (see Section 1.11; see Section 4.3).

Calibrating tempo under uncertainty. In fast-moving environments, review can lag behind deployment. PF's fast/slow posture treats reversibility, asymmetry, and dignity-of-awareness ceilings as signals that may justify widening review and raising the standard of evidence before proceeding (see Section 2.5; see Section 5.5).

These are not separate checklist items. They are facets of one stance: keep stakes, levers, and constraints visible while technical methods do their work.

6.5.2 Safety cases and other assurance artifacts as "bridge forms"

In some safety-critical domains, a **safety case / assurance case** is a structured argument, supported by evidence, intended to make safety reasoning explicit, reviewable, and revisable *[ISO/IEC/IEEE 2022]*. In AI governance, assurance-case methodologies have been explored as one way to make arguments about machine-learned components more structured and reviewable *[Paterson et al. 2025]*.

Safety cases are not the only bridge form. In less formal settings, incident reviews, design docs, model cards, risk registers, and monitoring plans can serve a similar function: keeping ethical and safety reasoning connected to concrete decisions (see Section 6.4). PF's decision record standard can also feed into assurance work by preserving assumptions, uncertainty, dissent, and revisit conditions in a form that later reviewers can evaluate (see Section 2.4).

When reading any such artifact through a PF lens, attention often turns to a few connected questions. Does the **scope of the claim** match the evidence provided, or is the argument broad while evidence is narrow (a form of "paper compatibility")? Where is **measurement weak or contested**, and what changes in response—narrower scope, stronger monitoring, improved measurement, pause/refusal pathways, stronger contestability, clearer repair commitments? How is **responsibility aligned** with Awareness × Power in practice, especially for those most affected? And is there a **path for revision** when conditions change, incentives distort reporting, or new harms appear (documentation without revision capacity can become fragile under real incentives *[Reason 1997]*)?

PF does not supply the technical evidence. Its role is to keep these ethical stakes and governance realities visible *inside* the argument, so reviewers can see what is being claimed, what remains uncertain, and who bears the consequences.

CC BY–NC 4.0

6.5.3 Limits and handoffs to safety engineering (technical and governance are entangled)

PF can inform the cultivated skill of Ethics—what we try to protect and how we justify tradeoffs—without supplying the technical content of safety engineering: hazard analysis methods, assurance techniques, security controls, formal verification, reliability analysis, red teaming, or monitoring design (see Section 4.6.3.3).

In practice, technical and governance challenges are rarely cleanly separable. A helpful way to use PF here is not to decide "which side" owns the problem, but to ask whether the response addresses both dimensions:

- **Technical dimension:** what evidence, tests, controls, and monitoring are adequate for the uncertainties and failure modes?
- **Governance dimension:** who can pause or change course, who can contest, what incentives distort reporting, and what remedies exist when harms occur?

PF helps teams pose these questions clearly; safety engineering supplies the methods and evidence to answer the technical parts of them.

6.5.4 Bridge to Part VII

Part VII turns to validation and adoption: how PF can be tested, taught, and iterated without turning it into a rulebook.

CC BY–NC 4.0

7 Validation and adoption roadmap

Part VII describes how Potentialism Framework (PF) v2.0 might be tested, adapted, and adopted without turning it into a badge, a certification regime, or a substitute for domain safety work. It treats adoption as a *practice problem* under incentives: a framework can be conceptually attractive yet still function in the world as a checklist, a shield, or a laundering device.

The posture here stays consistent with PF's broader stance: pilots are inquiry (not proof), documentation is a condition for contestability (not moral clearance), and revision is part of integrity (not embarrassment). Where stakes rise—especially near dignity-of-awareness concerns—the appropriate "success outcome" can include narrowing scope, slowing tempo, or handing off to stronger domain governance.

CC BY–NC 4.0

7.1 Status, scope, and interpretation

- **In scope:** how to read PF v2.0 as a proposed framework; what it is (and is not) claiming; how to interpret "adoption" in ways that preserve Responsibility rather than borrowing credibility.
- **Out of scope:** formal proof, certification, or universal scoring; claims that PF alone provides technical safety assurance (see Section 6.5.3).
- **Notes:** conceptual. This section clarifies posture so later validation talk does not smuggle in "PF is proven" as a background assumption.

PF v2.0 is offered as a proposed way of orienting attention: toward expressions-in-context, toward Compatibility as a multi-factor evaluation, and toward Responsibility as regulated responsibility (Awareness × Power (impact)). It is not a finished theory of everything, and it is not a compliance standard.

7.1.1 What PF is trying to be

PF can be used in at least three overlapping ways:

- **A conceptual lens** for noticing how context shapes expression, and how ethical evaluation shifts when awareness and power shift.
- **A governance lens** for making reasoning reconstructable and contestable—especially where incentives invite "ethics-washing" or "compatibility-washing."
- **A design and learning scaffold** for pilots: a way to surface assumptions, identify handoffs, and record revision triggers without pretending the pilot "proved" anything.

These uses can reinforce each other, but they also create different failure modes. A pilot scaffold can turn into a checklist; a governance lens can become a performance; a conceptual lens can be used as rhetoric. Part VII is designed to keep those risks visible.

7.1.2 What PF is not (and how misuse happens)

PF is not:

- **A technical safety case** or assurance method.
- **A replacement for domain governance** where PF is insufficient by itself (see Section 4.6).
- **A guarantee of Compatibility** (which depends on context, uncertainty, and contested values).

A common misuse pattern is to treat a PF artifact (a decision record, a score, a published summary) as evidence that harms were prevented or dignity was protected. Part VII treats

CC BY–NC 4.0

that pattern as a validation risk: the framework can be adopted while its core aims are bypassed.

7.1.3 What “validation” and “adoption” mean here

In PF’s posture, “validation” is less about proving a theory and more about testing whether PF improves governance capacity in practice: legibility, contestability, and repair. “Adoption” is not a single event; it is a pattern of use over time, under real incentives, with a revision trail that shows what changed and why.

7.1.4 Bridge to the next section

Section 7.2 clarifies a boundary that is easy to blur under pressure: the difference between analogy (helpful conceptual mapping) and validation (evidence that a practice works as intended).

CC BY–NC 4.0

7.2 Validation and analogy boundaries

- **In scope:** how to keep PF's conceptual commitments from being mistaken for empirical validation; how to use analogies without letting them do the work of evidence; what kinds of evidence and critique are relevant to PF pilots.
- **Out of scope:** claiming that PF's conceptual structure is "true" because it resembles other frameworks; importing domain evidence without checking fit.
- **Notes:** conceptual. The aim is to prevent "nice story" drift.

PF often uses analogies—between personal practice and institutions, between internal regulation and governance, between individual awareness and system-level awareness. Analogies can be useful for orientation, but they can also become shortcuts that look like validation.

A practical guardrail is to keep three questions separate in documentation and discussion:

1. **Is this an analogy or a claim?** If it is an analogy, what is it meant to illuminate—and what does it leave out?
2. **What would count as evidence here?** If the claim leans empirical (e.g., "this reduced harm," "this increased contestability"), name what observations would support or disconfirm it, and label the basis for any proxy inferences (see Section 2.4.3).
3. **What would count as "not working"?** If PF use is producing bureaucratic theater, suppressed contestation, or Responsibility diffusion, that is a form of negative evidence even if the narrative remains positive.

Validation, in this posture, is iterative and revisable: pilots produce traces that allow outsiders (or later insiders) to challenge the reasoning, not just the conclusions (see Section 2.4).

7.2.1 Bridge to the next section

Section 7.3 turns validation posture into practice design: what kinds of pilots to run, and how to sequence them so learning is real rather than cosmetic.

CC BY–NC 4.0

7.3 Pilot types and sequencing

- **In scope:** kinds of pilots that fit PF's posture; how to sequence pilots to reduce "checklist capture"; how to treat stopping, narrowing, or handoff as information rather than embarrassment.
- **Out of scope:** a universal maturity model or a required pilot order; claims that a pilot type is sufficient for all contexts.
- **Notes:** conceptual. The typology is meant to support choice, not impose a ladder.

PF pilots can be designed around different questions. These "types" are better read as families of inquiry than as boxes to tick.

7.3.1 Low-stakes legibility pilots

Aim: improve reconstructability without high-stakes enforcement.

Examples:

- writing decision records for representative cases and inviting critique of assumptions;
- labeling claim basis (conceptual vs empirical vs inferential) to reduce accidental overclaim;
- stress-testing the framework against known failure modes (washing, diffusion, retaliation risks).

Success here can look like clearer disagreement, sharper boundaries, and better handoffs—even if nothing "improves" yet at the outcome level.

7.3.2 Contestability pilots

Aim: test whether challenge can matter to a reasonable degree in the actual context (visibility, standing, remedy) (see Section 4.3).

Examples:

- creating a pathway for affected parties (or credible advocates) to contest a decision and observing whether contestation changes anything;
- testing whether disclosures create retaliation risk and adjusting publication posture accordingly.

CC BY–NC 4.0

7.3.3 Repair and follow-through pilots

Aim: test whether the system can respond "in time to matter" when harm signals appear (see Section 5.5).

Examples:

- rehearsing repair and restitution pathways;
- defining pause / escalation triggers and using them in at least one real or simulated case.

7.3.4 Boundary and handoff pilots

Aim: make explicit where PF is insufficient and where domain methods take over.

Examples:

- mapping where PF evaluation ends and technical safety assurance begins;
- documenting what PF can say (orientation, tradeoffs, contestability needs) and what it cannot responsibly certify.

7.3.5 Sequencing posture

In many contexts, starting with legibility pilots before high-stakes enforcement reduces incentives to perform. As stakes rise, shifting tempo toward "slow mode" (more time, broader review, more dissent capture) is often healthier than accelerating because attention is hot (see Section 2.5).

A pilot that stops early can be a strong learning outcome when (a) the reasons for stopping are made explicit, and (b) the stop reveals a boundary—capacity limits, dignity risks, missing contestability—rather than merely avoiding discomfort.

7.3.6 Bridge to the next section

Section 7.4 specifies the trace and publication posture that makes pilots shareable without turning them into propaganda or exposing low-power participants to retaliation.

CC BY–NC 4.0

7.4 Documentation and publication commitment

- **In scope:** what to document so PF work is reconstructable; how to publish in ways that support contestability while protecting dignity; how to report failures so pilots do not become borrowed credibility.
- **Out of scope:** a single universal disclosure policy; requirements to publish everything; incident response and safety assurance reporting standards.
- **Notes:** conceptual. The goal is calibrated transparency: enough trace for challenge to matter, without creating new harm pathways.

PF's claim is not "we are transparent." It is: we leave behind a trace that allows others to see what was evaluated, how uncertainty was handled, and what would change the judgment.

7.4.1 What to document so reasoning is reconstructable

Documentation is most useful when it distinguishes:

- **Expression-in-context:** what happened, under what constraints and incentives;
- **Evaluation posture:** what was treated as observation vs interpretation vs inference;
- **Compatibility considerations:** foreseeable effects on avoidable suffering/harm, dignity of awareness, and others' freedom to regulate their potentials within real constraints (see Section 1.7);
- **Decision and revisit conditions:** what was chosen, why, what remains uncertain, and what would trigger reconsideration (see Section 2.4).

7.4.2 What not to confuse with documentation

High-volume documentation can coexist with low contestability. If records are polished narratives that remove uncertainty and dissent, they can function as credibility artifacts rather than learning artifacts.

A useful internal question is: *could a critical outsider reconstruct the reasoning well enough to challenge it?* If not, "documentation happened" may not be the right success story.

7.4.3 Publication posture: calibrated transparency

PF's publication commitment is to publish enough to support contestability to a reasonable degree, while acknowledging that full disclosure can be coercive in asymmetric contexts.

In practice, publication choices often involve tradeoffs:

- **Who is protected by limiting disclosure?** (e.g., low-power participants)
- **Who is harmed by limiting disclosure?** (e.g., affected parties lacking visibility)

CC BY–NC 4.0

- **What alternate contestation path exists, if public disclosure is constrained?**

7.4.4 Reporting failures and negative results

To prevent "pilot as propaganda," it is often valuable to publish:

- failure modes encountered (washing incentives, retaliation risks, diffusion patterns);
- what was changed in response;
- what remained unresolved and why.

This turns "failure" into usable information and makes revision legible rather than rhetorical.

7.4.5 Dignity-of-awareness concerns and publication limits

Near dignity-of-awareness concerns, publication itself can become harm. PF's ceiling principle suggests resisting forms of exposure that would plausibly destroy or humiliate awareness beyond limits, even when disclosure would increase legibility (see Section 1.9).

When disclosure is constrained for dignity reasons, the task is not to pretend transparency was achieved, but to preserve as much contestability as constraints allow—through trusted reviewers, protected channels, or structured standing for challenge (see Section 4.3).

7.4.6 Bridge to the next section

Section 7.5 describes what "improvement" can mean for PF pilots, and what patterns suggest drift, misuse, or the need to pause and revise.

CC BY–NC 4.0

7.5 Success criteria and failure criteria

- **In scope:** what "success" and "failure" can mean for PF pilots; criteria for usefulness that do not collapse into "PF works / PF fails"; signals that a pilot is increasing legibility, contestability, and Responsibility rather than laundering it; revision signals and pause/escalation prompts calibrated to Awareness × Power (impact) and proximity to dignity-of-awareness concerns.
- **Out of scope:** certification, compliance scoring, or a universal pass/fail gate; claims that any one metric captures "Compatibility"; technical safety assurance and incident-response methods.
- **Notes:** conceptual. This section offers evaluation posture under real incentives. It aims to make improvement and drift inspectable and revisable, not to turn PF into a scoreboard.

Section 7.3 treated pilots as inquiry, and Section 7.4 described how that inquiry becomes shareable through trace. This section adds evaluation posture: what would count as improvement, and what would count as a reason to pause, revise, or narrow scope—so that "we ran a pilot" does not become borrowed credibility.

In PF's framing, pilots can be "successful" even when they reveal limits, trigger stops, or force revisions. "Success" here is not moral clearance. It is evidence that learning is becoming more legible and more contestable, and that this learning is connected to better outcomes for those affected—or, at minimum, to clearer identification of why outcomes cannot be improved in the present context.

7.5.1 What criteria are for (and what they are not)

Criteria can protect three things:

1. **Learning (against storytelling):** without criteria, outcomes can be narrated as success regardless of what changed, who carried cost, or whether challenge was possible.
2. **Responsibility (against diffusion):** criteria help keep answerability aligned with Awareness × Power (impact), rather than drifting toward scapegoating or "no one was responsible."
3. **Dignity of awareness (against instrumentalization):** criteria can surface when disclosure or enforcement pressures risk violating the ceiling principle.

CC BY–NC 4.0

These criteria are closer to evaluation questions than to a checklist. Different contexts will set different thresholds; different stakeholders will experience stakes differently. The point is to keep judgments discussable and revisable.

7.5.2 Success criteria: signals a pilot is becoming useful

Useful pilots tend to produce learning along several intertwined dimensions. None is sufficient alone, and a pilot may legitimately emphasize one dimension while naming limits on the others.

Legibility increased without false certainty. The reasoning trail makes it possible to reconstruct what was evaluated, what was inferred, and what would need to change for the judgment to flip. Legibility can also be weaponized (surveillance, control). A useful question is: *who gains from increased legibility, who could be harmed, and can those harmed contest how legibility is used?*

Contestability became more real (not just more visible). Some path exists for challenge to matter to a reasonable degree—visibility, standing, and remedy—even if full public disclosure is not possible (see Section 4.3).

Responsibility mapping improved (answerability stayed aligned with power). The trace shows where decisions could prevent or repair harm, and avoids collapsing "accountability" into humiliation or blame theater—especially near dignity-of-awareness concerns.

Follow-through occurred (repair, escalation, or stop happened in time). When harm signals appeared, repair or escalation pathways were used; revisit conditions were invoked rather than archived; stopping or narrowing happened when continuing would plausibly exceed the pilot's governance capacity (see Section 5.5). Stopping reads as a stronger success signal when the reasons are explicit and contestable, rather than retrospective rationalization.

The prototype changed (the framework learned). The pilot led to concrete revision: sharper boundaries, improved record formats, better contestability paths, or clearer handoffs.

7.5.3 Failure criteria: signals of drift, misuse, or insufficient fit

PF treats "failure" as first-class information. Still, some patterns are useful cues to pause, narrow scope, or revise before proceeding. These are not verdicts on participants; they are patterns to watch for.

CC BY–NC 4.0

- **Process substitution:** documentation volume rises while legibility and contestability do not.
- **Contestability collapse:** visibility exists without standing, or standing without remedy, so challenge becomes performative.
- **Responsibility diffusion or scapegoating:** obligations concentrate on low-power individuals while high-impact decision points remain unexamined.
- **Manufactured compatibility:** apparent "compatibility" is achieved by suppressing contestation, narrowing whose dignity counts, or redefining harms out of scope.
- **Category errors:** PF judgments are treated as technical safety assurance or used to bypass domain methods.

7.5.4 Calibrating criteria to stakes

Because Responsibility scales with Awareness and power (impact), PF's posture expects the justificatory burden to rise as stakes rise—without turning that into a fixed ladder.

Calibration benefits from explicitly asking: **stakes for whom?** A pilot that seems low-stakes to decision-makers may be high-stakes for affected parties whose dignity or freedom-to-regulate could be impacted.

As stakes and dignity proximity increase, evaluation typically benefits from some combination of:

- more explicit tradeoffs and uncertainties;
- stronger contestability paths;
- clearer revisit conditions;
- more practiced pause / escalation options.

When criteria conflict—for instance, transparency might harm low-power participants—it can be helpful to ask: who is protected by limiting disclosure, who might be harmed, and what alternative contestation paths exist. These questions do not resolve the conflict; they make the remainder visible.

7.5.5 Bridge to the next section

Section 7.6 turns evaluation signals into governance: how updates are proposed, decided, and recorded—so criteria lead to revision rather than rhetoric.

CC BY–NC 4.0

7.6 Revision mechanism and governance of updates

- **In scope:** how PF v2.0 is updated in a way that preserves legibility, contestability, and Responsibility; how proposals enter; how decisions are made and recorded; how disagreement is preserved; how versions are published and superseded with clear reader guidance.
- **Out of scope:** legal constitutions, certification regimes, or enforcement authorities; guarantees that the process prevents capture; technical safety assurance processes (see Section 6.5.3); governance design for a specific institution.
- **Notes:** conceptual. PF does not assume a single "maintainer class." In practice, updates may be proposed and adopted by whoever is using the framework; this section describes integrity-preserving postures that can be adapted to different settings.

PF's revisability posture is an explicit stance under uncertainty (see Section 3.1.5). If Responsibility scales with Awareness × Power (impact), then the governance of updates is one place that Responsibility either stays connected to practice—or drifts into rhetoric. It is also a place where Will and Ethics can show up in institutional form: regulating pace and scope (Will; see Section 1.5), and keeping compatibility-with-dignity in view under pressure (Ethics; see Section 1.6).

7.6.1 What "revision" is doing here

A revision mechanism is how learning becomes change: it translates pilots, critique, and disagreement into updated language and practices that others can inspect.

PF itself has potentials: it can be interpreted and expressed as a compass, as a checklist, or as a rhetorical shield. A revision mechanism is one way of shaping which of those potentials gets expressed in practice.

A revision mechanism is doing at least three jobs:

1. **Incorporating new information.** Does the process actually absorb what pilots and critiques revealed, or do incentives push it toward "nothing to see here"?
2. **Keeping challenge able to matter.** Contestability is structured capacity for challenge to matter to a reasonable degree (see Section 4.3).
3. **Keeping the dignity ceiling in view.** Where revisions touch high-stakes domains or dignity-of-awareness concerns, the update process may need to slow down, narrow scope, or defer to stronger governance (see Section 1.9; see Section 4.6).

CC BY–NC 4.0

"Measured" governance here means responsive to information rather than final, and willing to preserve disagreement rather than forcing consensus for narrative neatness.

7.6.2 What can change, and what is treated as "breaking"

PF updates can land in different layers:

- **Clarifications:** wording changes intended to reduce ambiguity without changing meaning.
- **Tooling / practice refinements:** updates to prompts, templates, record formats, or examples.
- **Boundary and handoff refinements:** sharper statements of where PF is insufficient by itself and what it hands off to domain methods.
- **Definition-lock-affecting changes:** changes that would alter the evaluative function of a core term—changing what counts as Will, Responsibility, Compatibility, or Dignity of awareness in ways that would lead to different judgments in otherwise identical cases (see Section 0.6).

Definition-lock-affecting changes are best treated as "breaking" changes in practice—often handled as a new major version—with explicit rationale, dissent capture, and a clear migration note for readers relying on prior wording. The aim is to prevent meaning changes from happening silently.

7.6.3 How proposals enter

PF does not require a single "intake" design. An open proposal pathway can work when it stays lightweight and when it does not treat templates as gates.

In practice, proposals are easier to evaluate when they clarify (in whatever format fits the context):

- where the change would land and what wording is proposed;
- what problem motivates it (confusion, misuse pattern, contradiction, missing boundary, newly discovered failure mode);
- what layer it touches (clarification vs practice vs boundary vs definition-lock-affecting);
- what reasoning or evidence supports it, and how claims are labeled when they lean empirical (see Section 2.4.3);
- what "better" and "worse" might look like after adopting it.

CC BY–NC 4.0

Where stakes are low, a short note can be enough. Where stakes are higher, proposals often benefit from "slow mode" tempo: more explicit tradeoffs, more contestability work, and a clearer revisit path (see Section 2.5).

7.6.4 How decisions are made without pretending neutrality

PF does not require a single governance topology. But it benefits from a consistent decision posture.

- **Calibrate review intensity to stakes—explicitly.** As Awareness × Power (impact) rises, it is often appropriate to raise justificatory burden and expand review. Calibration benefits from naming *whose stakes* are being considered, and who is missing from that assessment.
- **Use decision records for updates.** Accepted changes (and consequential rejections) are easier to evaluate later when accompanied by a decision record: what was chosen, why, what remains uncertain, and what would trigger reconsideration (see Section 2.4). Treat this as ethical memory, not bureaucracy.
- **Resist "artifact completion" substitution.** A full template is not the same as a good decision. Records exist to keep reasoning reconstructable, not to create a new compliance surface (see Section 2.7.6; see Section 6.4.6).

Where implementation is possible, one integrity-preserving asymmetry is to treat updates that would reduce contestability, narrow whose dignity is protected in practice, or increase harm risk as carrying a higher justificatory burden than updates that increase legibility and repair capacity—because the downside is often harder to unwind once normalized.

7.6.5 How disagreement is recorded (and kept alive)

PF treats disagreement as first-class information. Hiding disagreement can increase the risk of later re-litigation, narrative drift, and avoidable surprise—especially when decisions leave moral remainder (see Section 5.2).

When disagreement is substantive, records can preserve depth by noting (as space and risk allow):

- the strongest competing arguments (represented as fairly as possible);
- who is likely helped and harmed by each path (to a reasonable degree);
- what evidence or experience would change someone's mind;
- what remains contested after the decision;
- whether an alternate wording or "optional path" is being preserved.

CC BY–NC 4.0

When practical, preserve minority views as a dissent entry linked to the decision record, so later revisions do not have to rediscover the same objections (see Section 3.1.5). This is not an invitation to endless debate; it is a way to keep the framework honest about what it does and does not settle.

7.6.6 Publishing, versioning, and superseding guidance

A revision mechanism earns trust through trace:

- **Version releases where feasible,** with a change log readable by non-maintainers.
- In general, **try to avoid "silent patches"** to meaning-bearing language. If meaning changes, label it as such and point to the decision record that explains why.
- **Supersede with care.** When older language is retired, keep it accessible with a clear pointer to the replacement and a short note about what changed.
- **Keep reversibility available.** In open frameworks, "pausing distribution" may be limited once material is public. Still, having practiced ways to issue advisories, narrow claims, and clearly mark versions as superseded can help keep learning and Responsibility "in time to matter" (see Section 5.5).

This is also where PF's publication posture matters: not every internal deliberation is safe to publish. When disclosure would plausibly harm low-power participants or create coercive leverage, record the tradeoff and preserve as much contestability as constraints allow (see Section 4.3).

7.6.7 Bridge to appendices

The appendices can hold the concrete artifacts that make revision workable **without turning PF into a rulebook**: proposal examples, decision-record templates, disagreement-log examples, and the version history. Keeping these artifacts adjacent—but clearly subordinate—to the conceptual core helps maintain compass posture while still supporting real-world use.

CC BY–NC 4.0

Appendices

CC BY–NC 4.0

A. Extended glossary

- **In scope:** expanded clarifications, edge cases, and disambiguations behind the Mini glossary; brief PF-usage notes for recurring phrases used across Parts I–VII.
- **Out of scope:** replacing the Mini glossary as the authoritative definition anchor; settling broader philosophical disputes about these terms in general usage; legal definitions; new empirical claims about prevalence.
- **Notes:** conceptual. Where a term has a definition in Section 0.6, that definition remains the reference meaning. Material here is elaboration and disambiguation for use across PF; it does not alter, replace, or extend the locked meanings (see Section 0.6). This appendix is selective rather than exhaustive, and the appendices are meant to be usable modularly rather than only in sequence.

A.1 How to use this appendix

- **If you need the reference meaning:** use the Mini glossary definition (see Section 0.6).
- **If you need the application/disambiguation layer:** use the notes here as non-definitional guidance for how PF applies the locked terms across different contexts.
- **If a phrase is not listed here:** return to Part I for the core architecture or to the relevant Part for local usage; this appendix is a reference aid, not a complete dictionary.
- **If you are translating PF into another vocabulary:** treat the notes here as *functional targets* rather than word-for-word demands (see Part VI).

A.2 Core Mini glossary terms (definition-locked) + usage notes

Use note: The notes below clarify recurring PF usage. They are not alternate definitions.

➢ Potential

Definition (locked): Potential — a neutral underlying capacity of a system or being to generate patterns of experience or behavior when triggered. Sources of potentials may include biology/embodiment, learned habits, meanings/narratives/values, and extended systems/tools (including engineered components).

PF usage notes (non-definitional):

- "Neutral" here means: *not a moral verdict by itself*. Moral salience typically enters through **expressions-in-context** and their effects.

CC BY–NC 4.0

- PF may treat "potentials in play" as a working map of capacities and pressures (abilities, incentives, habits, tools, roles, constraints)—as a way to make what could be expressed more legible.
- A potential can be present without being expressed; PF often asks what (or who) is regulating expression, and what makes regulation feasible.

➢ Expression

Definition (locked): Expression — a concrete, situated manifestation of one or more potentials in a specific moment and context.

PF usage notes (non-definitional):

- PF often uses "expression" broadly: individual actions, system outputs, organizational decisions, and institutional behaviors—so long as the target is concrete and situated (see Parts I–II).
- "Situated" is the reminder that the same potential can express very differently under different contexts (power/impact, incentives, threat, time pressure).

➢ Context

Definition (locked): Context — the concrete configuration of conditions in which an expression takes place (e.g., relational field, power/impact, resources/constraints, incentives, norms/rules/instructions, awareness, and internal states).

PF usage notes (non-definitional):

- PF treats internal states as context factors insofar as they may shape which expressions are likely and how feasible regulation is in the moment.
- PF often highlights power/impact as a context feature that changes ethical stakes even when intent looks similar.
- "Norms/rules/instructions" are part of context too: they can enable regulation, or become laundering channels when treated as substitutes for judgment (see Part IV).

➢ Will

Definition (locked): Will — the trained capacity to regulate how, when, and how intensely potentials are expressed.

PF usage notes (non-definitional):

- **In application**, PF users sometimes map will-like regulation capacity to concrete control/constraint mechanisms: decision rights, interlocks, escalation pathways,

CC BY–NC 4.0

monitoring/rollback, and the practical ability to refrain (see Section 4.4; see Section 6.4). This is a translation move, not a change to the definition.
- Will can be constrained by context: threat, dependency, coercion, and lack of alternatives can reduce what "regulation" is realistically available (see Part IV).
- PF keeps **Will (regulation capacity)** distinct from **Ethics (regulative guidance)**: Will is the steering capacity; Ethics is the cultivated skill informing steering (see Part I).

➢ Ethics

Definition (locked): Ethics — the cultivated insight/skill of choosing expressions that are as mutually compatible as possible, while protecting dignity and reducing avoidable suffering.

PF usage notes (non-definitional):

- PF frames Ethics as trained judgment rather than a rule set, aiming to remain usable under uncertainty and conflict.
- "Mutually compatible as possible" is a compass orientation: it points toward tradeoffs and justificatory burden when full compatibility is not available (see Section 5.2).

➢ Compatibility

Definition (locked): Compatibility — evaluating expressions relative to their foreseeable effects on: (a) avoidable suffering/harm, (b) dignity of awareness, and (c) others' freedom to regulate their potentials within real constraints (to a reasonable degree).

PF usage notes (non-definitional):

- PF treats Compatibility as multi-lens (harm, dignity ceiling, freedom-to-regulate) rather than a single metric or score (see Part III).
- "Foreseeable" is context-sensitive and responsibility-sensitive: what counts as foreseeable typically expands with Awareness × Power (impact) and access to relevant evidence; affected parties' reports may also surface harms insiders miss (see Section 1.11; see Parts III–IV).
- "Avoidable" is comparative: it invites the question "avoidable relative to what feasible alternative, given constraints?"
- When applying the freedom-to-regulate lens, evaluators often consider practical exit options, contestability, retaliation risk, dependence, and information asymmetry—without treating any single item as definitive by itself (see Parts III–IV).

CC BY–NC 4.0

➢ Awareness

Definition (locked): Awareness — (operationally) the capacity to understand consequences, model self/others over time, and regulate expression.

PF usage notes (non-definitional):

- "Operationally" signals scope: PF uses Awareness as a practical handle for ethical evaluation without requiring a settled metaphysics of consciousness.
- PF expects contestation at the margins: what counts as "understanding," "self/other," or "over time" can vary across cultures and systems; PF routes this to translation humility (see Part VI).

➢ Dignity of awareness

Definition (locked): Dignity of awareness — dignity scales with awareness; it functions as a ceiling principle: no "systemic compatibility" can justify the destruction or humiliation of awareness beyond limits.

PF usage notes (non-definitional):

- PF often uses the ceiling principle as a guardrail against "ends justify means" stories that destroy, degrade, or humiliate awareness for system goals (see Parts IV–V).
- In practice, as options approach severe coercion, degradation, or humiliation, justificatory burden often rises; beyond limits, the ceiling principle functions as a constraint against treating "systemic compatibility" as sufficient justification (see Section 4.5; see Part V).

➢ Responsibility

Definition (locked): Responsibility — obligation proportional to awareness and power (impact) to use will in line with ethical insight, and to answer for effects.

PF usage notes (non-definitional):

- PF uses Responsibility in a scaling sense: "who has enough awareness and impact that they ought to be answerable, and in what ways?" (see Part VI).
- PF distinguishes Responsibility from **blame** (which can become punitive) and from **liability** (legal/institutional). These can overlap, but PF does not assume equivalence.

➢ Regulated responsibility (Awareness × Power)

Definition (locked): Regulated responsibility (Awareness × Power) — the higher the insight and the higher the impact, the more ethical weight is carried.

CC BY–NC 4.0

PF usage notes (non-definitional):

- PF treats this as a heuristic for allocating attention, safeguards, and answerability—not as a precise formula.
- In institutions, "power" often includes infrastructural forms of impact (control over resources, policy, enforcement, information, or platform constraints), not only interpersonal influence (see Part VI).

A.3 Supporting phrases in PF's core usage

➢ Expressions-in-context

A combined phrase used to keep **Expression** and **Context** together in evaluation. It is shorthand for PF's core unit of evaluation—not a separate concept—and points to what actually happens in a concrete situation, including downstream effects to a reasonable degree, rather than to what a system "is" in the abstract (see Part I; see Part VI).

➢ Harm / dignity / agency

A recurring PF shorthand for Compatibility's three lenses:

- **harm** → avoidable suffering/harm,
- **dignity** → dignity of awareness as a ceiling principle,
- **agency** *(used here only as shorthand for freedom-to-regulate)* → others' freedom to regulate their potentials within real constraints.

The shorthand is for readability; it does not replace the fuller definition (see Part I; see Section 0.6).

➢ Foreseeable effects

In PF usage, "foreseeable" is bounded by what can reasonably be anticipated given available Awareness, accessible evidence, and time—while generally expanding as Awareness × Power (impact) rises. The phrase is an orientation aid, not a mechanical rule, and "bounded" should not function as an evasion (see Parts III–IV).

➢ Real constraints

PF shorthand for constraints that materially limit feasible options (e.g., coercion, dependence, time pressure, resource scarcity, institutional incentives, retaliation risk, information asymmetry, governance limits) (see Parts IV–V). Naming constraints is not, by itself, a justification; PF asks which constraints are fixed vs. design defaults, who bears them, and what contestation/repair paths exist.

CC BY–NC 4.0

➢ “To a reasonable degree”

A qualifier marking that what counts as “reasonable” is context-dependent and often contested, especially by those experiencing the constraints. In PF, its role is to keep that contestation visible rather than to supply a loophole or fixed threshold (see Parts IV–VI).

➢ Ceiling principle

PF shorthand for constraints that cannot be traded away by “systemic” stories without a rising justificatory burden—especially for dignity-of-awareness harms (see Part V).

➢ Answer for effects

PF usage for the practice of maintaining answerability and contestability for outcomes—not only what was intended (see Section 1.10; see Parts II–V).

A.4 Operational scaffolds and artifacts (Part II vocabulary; not part of the locked core)

Scope note: The terms in A.4–A.8 are secondary PF vocabulary and working labels used in later Parts. They are not part of the locked core glossary in Section 0.6, and they do not carry the same definition-lock status.

➢ Artifact (PF usage)

A structured output (note, checklist, record, map) used to make reasoning more legible, contestable, and revisable. PF treats “artifact-as-proof” as a misuse risk when artifacts become certificates instead of supports (see Parts II–IV).

➢ Context snapshot

A lightweight capture of the relevant context for an evaluation or decision: who is affected, what constraints exist, what incentives shape behavior, what evidence is available, and what uncertainties remain (see Part II).

➢ Options set

A compact representation of realistic alternatives under consideration. In PF, its role is to keep tradeoffs visible and resist the drift toward treating the first available path as the only path (see Part II).

➢ Compatibility judgment

A recorded evaluative claim about an expression-in-context across the Compatibility lenses, including what is uncertain and what would change the judgment (see Parts II–III).

CC BY–NC 4.0

➢ Responsibility map

A record of where relevant Awareness and Power sit for a decision, who can change course, who can contest, and who bears consequences—used to reduce Responsibility diffusion (see Parts II–IV).

➢ Will-practice prompt

A reflective prompt meant to help translate ethical insight into a change in expression—especially in timing, scope, or intensity. PF treats these as examples, not a required routine (see Part II).

➢ Decision record

A trace of what was decided, why, what was assumed, what dissent existed, what constraints shaped feasibility, and what triggers revisiting the decision. PF treats decision records as supports for revision and answerability, not compliance theater (see Parts II and VI).

➢ Revisit conditions / triggers

Pre-committed conditions that would warrant re-opening an evaluation (new evidence, harm signals, changed context, incentive distortion, suppressed dissent, etc.) (see Parts II, IV, VII).

➢ Fast / slow calibration

PF language for adjusting posture with stakes, reversibility, and time pressure. "Fast" does not imply careless; "slow" does not imply paralysis. The point is to match deliberation depth to risk and lock-in (see Parts II and V).

A.5 Measurement and evidence vocabulary (Part III; not part of the locked core)

➢ Measurement posture

PF's stance that measurement supports learning, coordination, contestability, and answerability—but does not convert ethics into a score or produce moral clearance (see Part III).

➢ Proxy

A measurable signal used as indirect evidence for something that matters but is not directly measurable in full. PF expects proxy limits and asks what the proxy cannot tell you (see Part III).

CC BY–NC 4.0

➢ Goodharting / Campbell-style distortion (PF usage)

PF uses this as shorthand for a familiar risk pattern: when proxies become decision targets, they can distort attention and reporting—especially under incentives (see Part III).

➢ Audit substitution (PF usage)

Replacing lived reality with paperwork success ("we have the form"). PF names this as a misuse risk, especially under incentives for legibility and defensibility (see Parts III–IV).

➢ Selective visibility

A pattern where what is easy to count becomes more visible than what matters most ethically, especially when harms are hard to report or official channels are unsafe (see Part III).

➢ Power shielding

A pattern where those most affected have the least ability to generate recognized signals, shape metrics, or contest the official account (see Part III).

➢ Evidence (PF usage)

PF treats evidence as plural: outcome signals, process traces, behavior under conditions, and lived experience/testimony can all matter—none automatically settles Compatibility (see Part III).

➢ "Measured, but not settled"

PF shorthand for keeping measurement in its supportive role: metrics as evidence, not permission slips (see Part III).

A.6 Misuse-resistance vocabulary (Part IV; illustrative pattern-labels, not part of the locked core)

Note: These labels are pattern-language for incentive-shaped drift modes. They are not intended as moral indictments, assertions regarding motive, or empirical claims about prevalence. They are working labels used in the current PF draft and may be refined through later revision (see Part IV; see Part VII).

➢ Compatibility-washing

Using Compatibility language to legitimize an expression while avoiding the substantive questions PF is meant to keep visible: real constraints, lived impact, Responsibility scaling, contestability, and dignity-ceiling risk (see Part IV).

CC BY–NC 4.0

➢ Scope-squeezing

Narrowing the evaluation boundary so that key harms are treated as "out of scope," even when they are central to affected experience (see Part IV).

➢ Declared intention substitution

Treating stated intent as sufficient evidence of ethical alignment, while constraints, impacts, and lived experience are not addressed (see Part IV).

➢ Metric substitution / substitution-by-proxy

Treating a proxy (dashboard, certification, checklist) as if it is Compatibility, often erasing dignity-ceiling risks and freedom-to-regulate constraints (see Part IV).

➢ Dignity as ornament

A pattern where dignity language is rhetorically prominent but does not reliably constrain decisions or protect contestability in practice (see Part IV).

➢ Legitimation drift

A pattern where ethical language changes faster than practice, and new vocabulary is used to justify what would likely have happened anyway (see Part IV).

➢ Narrative capture

A pattern where the story of the action becomes more protected than the people affected by it, and dissent is reframed so scrutiny weakens rather than deepens (see Part IV).

➢ Moral credentialing

A pattern where past "goodness" is treated as evidence that present actions are compatible, reducing scrutiny exactly when scrutiny is needed (see Part IV).

➢ Washing-by-translation (PF usage)

Calling "implementation" or "translation" complete once artifacts exist (reports, dashboards, certifications), then treating those artifacts as evidence that ethical questions have been answered (see Parts IV and VI).

➢ Paper compatibility

A pattern where an argument appears ethically adequate because it documents what is easy to evidence, while leaving the highest-stakes expressions or constraints unaddressed (see Part VI).

CC BY–NC 4.0

➢ Boundary moves without answerability

Outsourcing, role-splitting, or relabeling moves that make Responsibility look diffuse without changing where Awareness or Power actually sit (see Part VI).

A.7 Conflict, thresholds, and repair vocabulary (Part V; not part of the locked core)

➢ Compatibilities collide

PF shorthand for cases where the Compatibility lenses pull in different directions, or where constraints make fully compatible options unavailable. PF treats collision as normal, not as proof of bad faith (see Part V).

➢ Tradeoff

A situation where no available option is fully compatible across lenses. PF tries to keep tradeoffs explicit, including who bears costs and what is being locked in (see Part V).

➢ Threshold (PF usage)

A qualitative posture shift, not a scoring rule: a point where the burden of proceeding "as is" rises enough to warrant a different stance (slower mode, narrower scope, stronger constraints, refusal of a channel, escalation) (see Part V).

➢ Moral remainder

The ethically relevant cost that can persist even after a careful attempt to choose the most compatible available expression—what remains "unpaid" by procedure or sincerity (see Part V).

➢ Repair / restitution

PF's forward-facing orientation to reducing harm, restoring agency, and making consequences answerable. Repair does not retroactively justify an incompatible action; it is part of carrying Responsibility after the fact (see Part V).

➢ Escalation / stopping rules

PF vocabulary for widening relevant awareness/authority when stakes are high or drift is likely, and for pausing/refusing channels when dignity-ceiling or irreversible harms are plausibly in play (see Parts V, IV, VII).

➢ "Moral cleaning"

PF shorthand for the misuse of repair language as if later restitution retroactively made an incompatible act ethically clean. PF rejects that move (see Part V).

CC BY–NC 4.0

A.8 Translation, supplementation, and handoffs (Part VI; not part of the locked core)

➢ Translation humility

PF posture that cross-cultural and cross-domain use is real translation work that can surface legitimate disagreement about meaning and legitimacy; PF aims for better dialogue and fewer hidden assumptions, not forced convergence (see Part VI).

➢ Translate by function

A heuristic: preserve what PF is doing (keeping harm, dignity ceilings, freedom-to-regulate, and Responsibility scaling visible) rather than demanding one-to-one word mapping (see Part VI).

➢ Supplementation posture

PF's stance that orientation is often not enough: domains may require evidence, enforceable governance, and (where relevant) technical assurance. PF can help keep ethical stakes visible inside those methods without replacing them (see Parts VI and IV).

➢ Technical assurance (PF usage)

PF uses this term as a pointer to established safety and engineering practices—testing, monitoring, assurance cases, verification, security controls, and related methods—that lie outside PF's conceptual scope but may be necessary for supplementation. PF also notes a risk: assurance artifacts can be misread as ethical clearance if they crowd out contestability and Responsibility alignment (see Section 6.5.3; see Parts VI and IV).

➢ Contestability

A structural condition for ethical learning in which affected parties can see, challenge, and potentially change a decision or its conditions. PF often treats contestability as depending on visibility, standing, and remedy (see Part VI).

➢ Ethical memory

PF usage for the role decision records can play over time: preserving assumptions, uncertainty, dissent, and revisit conditions so later reviewers can understand what was believed and why (see Part VI).

➢ Mismatch detection

PF usage for one way responsibility mapping can help notice configurations like power without awareness, awareness without power, or affected people without standing—especially in complex socio-technical settings (see Section 6.4.2).

CC BY–NC 4.0

B. Printable compatibility checklist

- **In scope:** a one-page, printable prompt scaffold for quick Compatibility triage in real situations; a compact aid for keeping PF's three lenses visible under time pressure.
- **Out of scope:** a scoring tool; a certification device; a substitute for the fuller decision protocol; legal or technical assurance guidance.
- **Notes:** conceptual. This appendix condenses the Compatibility checklist posture in Part II into a printable format, while keeping misuse-resistance in view. It is best used as a prompt scaffold, not as proof that a decision is ethically settled; checked boxes record what was considered in a fast pass, not clearance or verdict.

B.1 How to use this checklist

This checklist is for moments when a fast pass is more realistic than a full protocol. It keeps the most important PF questions visible when time, attention, or coordination bandwidth is limited.

It can function as:

- a quick scan before acting,
- a shared prompt for team discussion and legibility,
- or a brief trace of why something was paused, narrowed, changed, or escalated.

Use it as one possible entry point under constraint, not as a required procedure for every case. If stakes are high, harms may be severe, dignity of awareness is plausibly at risk, reversibility is low, or disagreement is meaningful, the fuller protocol in Part II is often a better fit.

B.2 Printable checklist (core)

CC BY–NC 4.0

PF Compatibility checklist (prompt scaffold)

These are prompts for reflection, not items to verify.

Expression / decision under review: _____

Date / context: _____

Role(s) present (optional): _____

1) Name the expression-in-context

☐ What is the specific **expression** being considered or already happening?

☐ What is the decision boundary *right now* (what can still change)?

☐ What context features matter most here (power/impact, incentives, constraints, urgency, norms/instructions, awareness gaps)?

2) Fast three-lens scan (Compatibility)

A) Avoidable suffering/harm

☐ What avoidable suffering/harm might be foreseeable given our role, information access, and time—especially where impact is high?

☐ Who is most likely to bear the cost?

☐ What harms could be reduced by changing scope, timing, intensity, or safeguards?

B) Dignity of awareness (ceiling principle)

☐ Is there a plausible risk of humiliation or destruction of awareness beyond limits (ceiling principle)?

☐ If awareness is present—or plausibly in play—are we risking treating it as mere instrument under coercion, captivity, forced dependence, or severe asymmetry?

☐ Is a less dignity-threatening path being seriously considered?

C) Freedom-to-regulate (within real constraints)

☐ Whose freedom to regulate their potentials within real constraints is being reduced (to a reasonable degree)?

☐ Which limits look fixed for now, and which look more like defaults, incentives, or design choices that could be changed?

☐ Are we narrowing options temporarily and revisably, or locking them in?

CC BY–NC 4.0

3) Responsibility scan (Awareness × Power)

☐ Where do awareness and power/impact concentrate?

☐ Where do formal authority and practical power/impact differ?

☐ Who can still change course (pause, narrow scope, reroute, refuse, escalate)?

☐ Where might responsibility diffuse or drift toward low-power parties?

☐ If this later needs explanation, who will likely have to answer for effects?

4) Reversibility and posture shift

☐ If we are wrong, how hard is this to undo?

☐ Is this setting a precedent, locking in a path, or reducing future contestability?

☐ Is there a reason to slow down, narrow scope, or escalate/widen review?

5) Options (less harmful / more revisable)

☐ What variants could reduce harm while still serving the core purpose?

☐ What pause/delay option would buy learning, consent, or coordination?

☐ What smaller-scope or more reversible option is available?

☐ Are we treating the first available path as the only path?

6) Next-step orientation (non-binding)

Which of these, if any, seems warranted for now?

☐ Proceed as framed

☐ Proceed with changes / safeguards

☐ Pause / slow down

☐ Escalate / widen review

☐ Do not proceed via this channel (for now)

☐ Other: _____

CC BY–NC 4.0

Why (1–3 lines):

Brief notes (optional): tradeoffs, uncertainties, or revisit triggers

Back-of-page guardrails

- A checked box is not a verdict.
- “Constraint” is not, by itself, a justification.
- If the most affected parties are invisible here, the picture may be incomplete.
- If this is being used only to justify a decision already made, it has stopped being useful.

When this checklist is not enough

A fuller pass is often worth considering when impact is large or hard to reverse, disagreement is meaningful, power is concentrated while others bear the costs, or empirical assumptions are doing major work. In those cases, PF often benefits from the fuller decision protocol, responsibility mapping, and decision-record scaffolds in Part II.

CC BY–NC 4.0

C. Printable decision record template

- **In scope:** a fill-in template for recording decisions in a way that keeps reasoning reconstructable, reviewable, and revisable; a shared shape for documenting context, options, judgment, uncertainty, and revisit conditions.
- **Out of scope:** legal compliance documentation; a mandatory workflow; proof that a decision was ethically correct; a substitute for governance, technical assurance, or fuller domain evidence.
- **Notes:** conceptual. This appendix turns the decision-record posture in Part II into a printable scaffold. It is best used as a container for traceable reasoning, not as evidence that ethical questions are settled. A completed record shows what was considered at the time; it does not constitute clearance, verdict, or proof of Compatibility.

C.1 How to use this template

This template is often useful when a decision is likely to matter later—for example when stakes are high, precedent may be set, reversibility is low, disagreement is meaningful, or safeguards and monitoring are part of the rationale for proceeding.

It can function as:

- a compact record for team decisions that may need later review,
- a trace note for why a path was chosen, deferred, narrowed, or escalated,
- or a shared container for keeping assumptions, tradeoffs, and revisit conditions visible over time.

It is not meant to create routine bureaucracy. Use it selectively and scale detail to stakes. In lower-stakes settings, only part of this template may be worth using. In higher-stakes settings, more detail can help the reasoning remain reviewable and revisable.

Safest useful specificity: record enough detail to support review and repair, while protecting sensitive information where disclosure would create avoidable harm.

CC BY–NC 4.0

C.2 Printable PF decision record template

Short label / case name: _____

Date / time: _____

Decision-maker(s) / role(s): _____

Related case / project / system boundary (optional): _____

1) Basis for key claims

Briefly indicate the nature of the claims to help orient future review. *(Transparency cue, not a rating or clearance.)*

Basis for key claims (optional): conceptual / mixed / evidence-leaning / other

_____

_____

Notes on basis (optional):

_____

_____

2) Expression + decision boundary

- **Expression-in-context under review** (the concrete action / output / policy / omission being chosen):

  _____

- **Decision boundary** (what is in scope now, and what is not):

  _____

- **Time horizon** (one-time act / repeated practice / precedent-setting / ongoing):

  _____

CC BY–NC 4.0

3) Context snapshot

Consider noting the context features most likely to change the judgment (at the safest useful level of specificity).

- **Power / impact** (scale, reach, reversibility):

- **Incentives / pressures** (what is rewarded, punished, rushed, hidden):

- **Constraints** (resources, time, coordination limits, safety limits):

- **Norms / rules / instructions** (formal and informal pressures):

- **Awareness** (who can understand likely consequences well enough to steer or contest, and who bears risk without that visibility):

- **Internal states** (e.g., urgency, fatigue, fear, overload—only if materially relevant, structurally important, and safe to record):

- **Record-pressure note (optional):** what may be shaping the record itself?

CC BY–NC 4.0

4) Options set

List the realistic options considered, including smaller-scope or more reversible paths where relevant.

Option A:

Benefits / risks / constraints:

Option B:

Benefits / risks / constraints:

Option C / pause / defer / narrow-scope option (if relevant):

Benefits / risks / constraints:

CC BY–NC 4.0

5) Compatibility summary (three lenses, as reasons)

(Use these lenses as reasons for judgment, not as a score or verdict. They support summary; they do not exhaust the judgment.)

A) Avoidable suffering / harm

- Main foreseeable harms:

- What looks avoidable through changes in scope, timing, intensity, safeguards, or alternative path:

B) Dignity of awareness

- Any dignity-of-awareness (ceiling principle) concern, or uncertainty about whether such a concern is plausibly in play:

- If restraint, forced dependence, severe asymmetry, coercion, or humiliation risk is relevant:

C) Others' freedom to regulate their potentials within real constraints (to a reasonable degree)

- Whose freedom-to-regulate is reduced, and how:

- Which limits look fixed, and which look more like defaults, incentives, or design choices:

Optional short summary (non-binding):

If a brief summary helps later review, record it here in your own words rather than as a score or verdict.

CC BY–NC 4.0

6) Responsibility map (Awareness × Power)

(For answerability aligned with Awareness × Power (impact), not blame assignment or fixed duty allocation; see Section 1.11)

- Where do awareness and power/impact concentrate?

- Who can still change course, slow down, narrow scope, or escalate?

- Who is most affected with least voice / exit / contestability?

- If explanation or repair becomes necessary later, who will need to answer for effects?

---

7) Uncertainties, assumptions, and disagreement

- **Key assumptions doing major work:**

- **Top uncertainties / what could flip the judgment:**

- **Dissent / disagreement / representation gaps (if present):**

---

8) Safeguards, monitoring, and revisit triggers

- **Safeguards / boundaries / checks:**

- **Monitoring targets / signals to watch:**

- **Revisit trigger(s):** “reassess if X changes”

CC BY–NC 4.0

9) Chosen path

Describe the path taken in your own words. *(Examples: proceed as framed; proceed with changes or safeguards; pause / slow down; escalate / widen review; defer pending more evidence or coordination; do not proceed via this channel for now.)*

Why this path was chosen (2–6 lines):

10) Minimal answerability note

If someone most affected asked later, “Why was this done this way?”, what would they likely need to know to understand how this decision was reached, what constraints shaped it, and what remained uncertain or contested?

CC BY–NC 4.0

C.3 Pocket-note PF decision record template

When only a brief record is realistic, this shorter version can still preserve traceability:

Expression: _____

Context / why (1–2 lines): _____

_____

Top uncertainty: _____

_____

Revisit trigger: _____

_____

- This record is a container, not a verdict.
- A completed record does not, by itself, prevent responsibility drift, narrative capture, or paperwork theater.
- "Constraint" is not, by itself, a justification.
- If the record contains only the decision-makers' view, it may miss the highest-stakes realities.
- A clean narrative can still hide uncertainty, offloaded costs, or responsibility drift.
- If the record is mainly serving defensibility after the fact, it may no longer be serving learning.

When this template is not enough

A fuller pass is often worth considering when dignity-sensitive harms are plausibly in play, irreversibility is high, disagreement remains unresolved, empirical assumptions are carrying major weight, or many affected parties have limited voice, exit, or recourse. In those cases, PF often benefits from the fuller decision protocol, responsibility mapping, and validation/revision scaffolds in Parts II, VI, and VII.

CC BY–NC 4.0

D. Training curriculum materials

- **In scope:** exercises, examples, and learning-progression artifacts that help people use PF's scaffolds with more fluency over time; materials that support practice across novice, practitioner, and auditor postures.
- **Out of scope:** a mandatory curriculum; a certification regime; a single pedagogy for all settings; proof that training alone prevents misuse; a substitute for domain expertise, governance, or technical assurance.
- **Notes:** conceptual. This appendix translates Part II's training-pathway posture into reusable learning materials. It is intended to support judgment under real constraints, not to create a new compliance surface. The materials below are optional practices and examples, not a required sequence, official pedagogy, or credential path.

D.1 How to use these materials

These materials accompany PF's practical scaffolds, especially the compatibility checklist, decision-record template, responsibility mapping, and related prompts. The aim is not "coverage" for its own sake. The aim is to make a few recurring capacities more available when situations are pressured, contested, or easy to rationalize.

This appendix can function as:

- a menu of exercises for workshops, team learning, or self-study,
- a way to calibrate progression across novice, practitioner, and auditor postures,
- or a source of reusable drills that keep PF from collapsing into slogans, paperwork, or one-lens moralizing.

No sequence here is mandatory, and no exercise counts as proof of fluency. In many settings, a small subset may be enough. A short exercise repeated well can be more useful than a large curriculum performed once and forgotten.

D.2 Training posture: what this appendix is trying to cultivate

PF training is often less about memorizing answers than about improving a few repeatable capacities in context:

- noticing **expression** rather than drifting into essence-labeling, and seeing **context** with enough resolution to keep judgment situated,
- seeing **context** with enough resolution that "constraints" do not become automatic excuses,

CC BY–NC 4.0

- using Compatibility's three lenses—**avoidable suffering/harm; dignity of awareness (ceiling principle); and others' freedom to regulate their potentials within real constraints (to a reasonable degree)**—without collapsing them into one score or one feeling,
- connecting **Will** to real regulation moves (timing, scope, intensity, channel),
- keeping **Responsibility** aligned with Awareness × Power rather than diffusing it,
- leaving enough trace that later review can reconstruct what mattered,
- and recognizing when an artifact is helping real judgment versus simulating it.

These capacities do not necessarily develop in a straight line. People may move back and forth across postures as domain complexity, time pressure, and power dynamics change.

D.3 Learning progression map (non-linear by design)

These are descriptive postures of fluency and review, not sequential ranks or certifications.

These postures are orientation aids, not ranks or certifications. People may move among them across cases and settings, and different postures may be relevant at the same time.

D.3.1 Common novice orientation

Primary aim: basic PF literacy plus one or two reliable moves under constraint.

What often matters most here:

- identifying expression vs. context,
- naming the three Compatibility lenses without collapsing them,
- making one small will-regulation move,
- and leaving a minimal trace when a decision may matter later.

Typical exercise fit:

- short case scans,
- pair drills,
- one-minute or five-minute prompts,
- checklist practice,
- and "what changed the judgment?" reflection.

D.3.2 Common practitioner orientation

Primary aim: reliable use of PF under disagreement, higher stakes, and organizational friction.

CC BY–NC 4.0

What often matters most here:

- distinguishing material constraints from design/default choices,
- widening the option set where possible,
- separating what is known, inferred, and uncertain,
- making tradeoffs visible across the three lenses,
- and documenting without turning documentation into ritual.

Typical exercise fit:

- fuller case comparison,
- decision-record drafting,
- responsibility mapping,
- repair-oriented reflection,
- and scenario revision after new evidence or dissent.

D.3.3 Common auditor orientation

Primary aim: review decisions and systems for responsibility alignment, rationalization risk, and dignity-of-awareness pressure—without turning review into punitive enforcement or a status contest.

What often matters most here:

- reading records skeptically but fairly,
- spotting omissions, softened claims, or paperwork substitution,
- comparing formal roles to actual Awareness × Power,
- and checking whether revisit triggers, contestability, and repair pathways are likely to matter in practice.

Typical exercise fit:

- red-team review of records,
- artifact critique,
- discrepancy mapping,
- dissent reconstruction,
- and “what would count against this conclusion?” drills.

D.4 Design cues for training without compliance drift

When these materials are used, a few cues often help keep the training aligned with PF’s posture:

CC BY–NC 4.0

- **Treat artifacts as scaffolds for reasoning, not as scoring devices.**
  The checklist, decision record, and responsibility map are containers for reasoning, not proof that a decision is ethically settled.
- **Train on tradeoffs, not only on clean cases.**
  A recurring risk is learning performance rather than judgment when cases have an obvious "good answer."
- **Keep disagreement legible.**
  A useful session can preserve strong competing readings rather than forcing premature consensus.
- **Calibrate depth to stakes.**
  Fast-mode exercises matter, and so do exercises that slow people down enough to see what urgency hides.
- **Make contestability possible inside the training itself.**
  When participants cannot question the framing (to a reasonable degree), training can reproduce the same failures PF is trying to reduce.
- **Treat the training artifact as an expression in context.**
  Who designed the exercise, what it omits, what it rewards, and whose voice is missing can be ethically relevant.

D.5 Core exercise set

The prompt shapes below are starting points for reflection, not tests with correct answers.

D.5.1 Architecture drill: expression, context, and potential

Useful when: people are new to PF, or when discussion keeps drifting into identity labels, virtue language, or abstract judgments.

Prompt shape:

- Consider the **expression** actually under review.
- Which **potentials** might be active?
- Which **context** features could most change the meaning or effects?
- What would be distorted if we judged the actor's "essence" instead of the expression-in-context?

Useful output:

- a short expression/context distinction,
- one or two plausible potentials in play,
- and one note on what becomes more legible once the case is described concretely.

CC BY–NC 4.0

Common drift to watch:

- re-labeling the person instead of the expression,
- treating intentions as the whole case,
- or treating context as automatic exoneration.

D.5.2 Three-lens scan drill

Useful when: learners need practice holding harm, dignity of awareness, and freedom-to-regulate together without collapsing them.

Prompt shape:

- What avoidable suffering/harm seems most foreseeable here?
- Is any dignity-of-awareness ceiling concern plausibly in play?
- Whose freedom to regulate their potentials is being reduced, and under what real constraints?
- Which lens is easiest to see here, and which is easiest to ignore?

Useful output:

- one sentence per lens,
- one uncertainty,
- and one note about where the lenses pull in different directions.

Common drift to watch:

- one-lens capture,
- turning the third lens into vague “agency” talk (others’ freedom to regulate potentials),
- or using the word “constraint” without specifying who bears it and why.

D.5.3 Will-regulation drill

Useful when: the challenge is not “spot the issue” but “what concrete modulation is available now?”

Prompt shape:

- What potential seems to be driving this expression?
- What change in timing, scope, intensity, or channel would reduce avoidable harm?
- What pause, check, or reroute might make the expression more compatible without pretending the context is ideal?
- What would still remain difficult even after that move?

CC BY–NC 4.0

Useful output:

- one small regulation move,
- one more structural change that would help later,
- and one note about what context is narrowing Will in this case.

Common drift to watch:

- imagining perfect self-regulation,
- moralizing emotions instead of regulating expression,
- or proposing changes that are not actually feasible in the context.

D.5.4 Options-set drill

Useful when: the group is stuck between a default action and refusal, or when false dilemmas are shaping the discussion.

Prompt shape:

- What are at least three viable options within real constraints?
- What smaller-scope, slower, or more reversible option exists?
- Which option is easiest to defend on paper but weakest in lived effect?
- What would need to stay true for each option to remain acceptable?

Useful output:

- a short options set,
- one tradeoff note per option,
- and one revisit trigger that would reopen the judgment.

Common drift to watch:

- treating the first available path as the only path,
- confusing speed with necessity,
- or generating "options" that are not viable within the actors' real constraints.

D.5.5 Decision-record drill

Useful when: learners need practice preserving reasoning without drifting into paperwork theater.

Prompt shape:

- Consider the expression under review.
- Which context features most changed the judgment?
- What are the main reasons across the three Compatibility lenses?

CC BY–NC 4.0

- What assumptions are doing major work?
- What would trigger a revisit?

Useful output:

- a compact decision record using Appendix C,
- one dissent or uncertainty note,
- and one reflection on what the record still fails to capture.

Common drift to watch:

- writing polished narratives that hide uncertainty,
- omitting the path not chosen,
- or making the record look complete while leaving the highest-stakes tradeoff implicit.

D.5.6 Responsibility-map drill

Useful when: handoffs, automation, committee structure, or institutional layering make answerability hard to see.

Prompt shape:

- Where do Awareness and Power concentrate?
- Who can still change course in time to matter?
- Who bears the greatest costs with the least visibility, standing, or exit?
- Where is responsibility likely to drift if nothing is made explicit?

Useful output:

- a simple map of actors, affected parties, and influence paths,
- one note on power without awareness vs. awareness without power,
- and one suggestion for reducing responsibility diffusion.

Common drift to watch:

- turning the map into blame assignment,
- equating formal authority with practical power,
- or treating the current structure as fixed when some of it is design-dependent.

D.5.7 Misuse-resistance drill

Useful when: a team already has artifacts but may be sliding into substitution, washing, or rhetorical safety.

CC BY–NC 4.0

Prompt shape:

- What in this case could be doing the work of artifact-as-proof?
- Are we mistaking a record, metric, or process for ethical judgment?
- What tradeoff remains real here, even if the documentation looks clean?
- In what way might this artifact protect status more than improve answerability?

Useful output:

- one suspected drift pattern,
- one concrete revision to the artifact or process,
- and one question the team has been avoiding.

Common drift to watch:

- using the drift language as a purity test,
- assuming bad motives when structural explanation is more useful,
- or forgetting that the review posture itself is also an expression in context.

D.5.8 Dissent and disagreement drill

Useful when: the group is converging too quickly, or when later review would benefit from knowing what remained contested.

Prompt shape:

- What is the strongest competing reading of this case?
- What evidence or experience would change your mind?
- Who is likely helped and harmed by each path, to a reasonable degree?
- What should remain visible in the record even if a decision must still be made?

Useful output:

- one fair statement of dissent,
- one uncertainty that remains alive,
- and one note about whether the disagreement is conceptual, empirical, or mixed.

Common drift to watch:

- forcing consensus for neatness,
- preserving only weak dissent,
- or treating disagreement as failure rather than information.

CC BY–NC 4.0

D.6 Example session shapes

These are examples that can be adapted to available time and goals, not preferred or approved formats.

D.6.1 Short session (20–30 minutes)

- 3-minute orientation: remind participants that PF is a compass, not a scorecard.
- 7-minute architecture drill on a compact case.
- 10-minute three-lens scan or options-set drill.
- 5-minute reflection: what became more visible, and what still feels under-described?

D.6.2 Working session (45–60 minutes)

- Brief orientation.
- Compatibility checklist pass using Appendix B.
- Small-group decision-record draft using Appendix C.
- Whole-group comparison: where did judgments diverge, and why?
- Closing note on what would trigger revisit.

D.6.3 Review session (60–90 minutes)

- Read an existing record or case.
- Run responsibility-map and misuse-resistance drills.
- Add dissent and revisit entries.
- End with one process change that would improve answerability next time.

In some settings, one well-chosen exercise may be the better fit.

D.7 Learning progression artifacts

Some facilitators find lightweight artifacts like these useful for keeping learning visible without creating a credential ladder:

- **Practice log:** a short record of cases used, what posture was being trained, and what drift patterns showed up.
- **Reflection note:** one page on “what I missed first” / “what changed my judgment” / “what I still do not know.”
- **Drift-pattern tracker:** a recurring note on patterns such as sloganizing, one-lens capture, process substitution, committee fog, or blame theater.
- **Calibration set:** two or three cases that are periodically re-read to notice whether interpretations are becoming more careful, narrower, or more ritualized.

CC BY–NC 4.0

- **Facilitator note:** what the exercise surfaced, what stayed suppressed, and what would make the next round safer or more honest.

These artifacts are optional supports for ethical memory and revisability, not tools for ranking people.

D.8 Facilitator cautions

In training, it can help to remain aware that:

- a smooth workshop may still hide the absence of contestability,
- a polished answer may still be weak reasoning,
- repeated use of the same case type can narrow moral imagination,
- low-power participants may notice harms the training design itself makes hard to say,
- in some settings, confidentiality and safest-useful-specificity matter more than completeness,
- and a curriculum that cannot itself be revised may quietly train rigidity rather than judgment.

D.9 When these materials are not enough

Even well-run exercises do not replace governance, technical assurance, contestation, or repair. These materials may not be enough, for example, when:

- dignity-of-awareness risks are plausibly severe,
- reversibility is low,
- empirical or technical assumptions are carrying most of the conclusion,
- many affected parties have limited voice, exit, or recourse,
- or the organization is under incentives likely to turn training into performance.

In those cases, PF often benefits from the fuller operational scaffolds in Part II, the misuse-resistance posture in Part IV, the translation/handoff posture in Part VI, and the revision/validation commitments in Part VII.

CC BY–NC 4.0

E. Case library template

- **In scope:** a shared template for documenting cases, disagreements, outcomes, and revisions over time; a way to preserve trace across examples without flattening uncertainty, dissent, or context.
- **Out of scope:** a mandatory archive format; a leaderboard of "good" and "bad" cases; proof that a documented case was handled well; a substitute for the decision record, checklist, or fuller domain-specific documentation.
- **Notes:** conceptual. This appendix offers a reusable container for accumulation across cases. It aims to support reconstructability, contestability, and revision—not to create a case repository that functions as status display or borrowed credibility (see Section 2.4; see Part III; see Part VII). A completed entry preserves what was visible, contested, and recorded at the time; it does not constitute clearance, proof of correct application, or a closed moral verdict.

E.1 How to use this template

A case library can help when a framework needs more than memory. It can preserve how judgments were made, what remained contested, what later happened, and what changed after review.

This template can function as:

- a reusable shape for recording cases across teams or time,
- a way to compare similar cases without pretending they are identical,
- or a learning container for disagreement, follow-through, and revision.

Not every case entry needs every field. In lower-stakes or privacy-sensitive settings, a compact version may be enough. In higher-stakes, contested, precedent-setting, or dignity-sensitive cases, more detail can help later reviewers reconstruct the reasoning rather than inherit only a polished narrative.

Use the template selectively and scale detail to context. Omissions do not automatically make an entry weak; they may reflect safety, privacy, or practical limits. When fields are omitted due to constraint or safety, it can help to note what was omitted and why (at the safest useful level of specificity), so "missingness" does not silently turn into certainty (see Section 2.4; see Part III).

E.2 What this appendix is trying to preserve

A PF case library is meant to preserve more than the final answer. It can help keep visible:

- the **expression-in-context** that was evaluated,

CC BY–NC 4.0

- the **options set** that was or was not considered,
- the **Compatibility reasoning** across avoidable suffering/harm, dignity of awareness, and others' freedom to regulate their potentials within real constraints (to a reasonable degree),
- the **Responsibility** picture: where Awareness and power/impact concentrated,
- the **disagreement** that remained alive,
- the **outcomes and follow-through** that later mattered,
- and the **revision trail** showing what changed and why.

The point is not completeness for its own sake. The point is to leave enough trace that later readers do not have to reconstruct everything from status, hindsight, or a single authorized story.

E.3 Case selection note (to keep selection effects visible)

In the PF orientation, a case library is often more useful when it does not preserve only low-conflict, well-documented, or reputationally comfortable examples. It can help to note briefly why a case was included, so selection effects remain more visible rather than disappearing behind a tidy archive.

Useful selection reasons may include:

- the case is high-impact, hard to reverse, or precedent-setting,
- disagreement was meaningful,
- power was concentrated while others bore cost,
- harms were delayed, diffuse, or difficult to count,
- the case exposed artifact-as-proof, washing, or responsibility drift,
- or the case later required revision, repair, narrowing, or superseding.

This note is not a badge, and it does not by itself prevent selective archiving. It is a reminder that what gets archived is itself an expression in context.

CC BY–NC 4.0

E.4 PF Case library template (shared shape / prompt scaffold)

Use the fields below as modular prompts, not as a requirement to complete everything.

Case ID / short label: _____

Version / entry date: _____

Prepared by / role(s): _____
(for traceability, not blame)

Status:

☐ draft ☐ active entry ☐ revised ☐ superseded ☐ archived

Confidentiality / publication posture (optional): _____

Why this case is included: _____

1) Case snapshot

- **Expression under review** (the concrete action / output / policy / omission):
  _____
- **Decision boundary** (what was in scope at the time, and what was deferred or handled elsewhere):
  _____
- **Time horizon** (one-time act / repeated practice / precedent-setting / ongoing):
  _____
- **Setting / domain / system boundary (optional):**
  _____

CC BY–NC 4.0

2) Context features that most changed the case

- **Power / impact:**

- **Incentives / pressures:**

- **Constraints** (including whether some looked fixed vs. design-dependent):

- **Norms / rules / instructions:**

- **Awareness distribution** (who could understand likely consequences well enough to steer or contest, and who bore risk without that visibility):

- **Relevant context not available at the time (if known):**

  ---

3) Basis for key claims

(Basis-for-claims labeling; see Section 2.4.3. Cues for scrutiny, not ratings or clearance.)

☐ **Conceptual** (primarily normative / interpretive judgment)

☐ **Mixed** (concept + some evidence)

☐ **Evidence-leaning** (primarily external data / analysis)

What evidence, experience, or reasoning was most relied on:

What was missing, excluded, or weakly supported:

CC BY–NC 4.0

4) Options set and chosen path

- **Options actually considered:**

- **Least-harm / reduced-scope / more reversible variant (if any):**

- **Pause / defer / narrow-scope path (if considered):**

- **Chosen path:**

- **Why this path was chosen at the time (2–5 lines):**

5) Compatibility summary (three lenses, as reasons)

(See Section 1.7. Use these lenses as reasons, not as a score or verdict.)

A) Avoidable suffering / harm

- Main foreseeable harms:

- What was treated as avoidable, and relative to which feasible alternatives:

B) Dignity of awareness (ceiling principle)

- Any dignity-of-awareness concern, or uncertainty about whether such a concern was plausibly in play:

CC BY–NC 4.0

- Where restraint, humiliation, coercion, severe asymmetry, dependence, or destruction risk mattered:

C) Others' freedom to regulate their potentials within real constraints (to a reasonable degree)

- Whose freedom-to-regulate was reduced, and how:

- Which limits looked like real constraints, and which looked more like defaults, incentives, or design choices:

Optional short judgment label (non-binding):

6) Responsibility map summary

(For answerability aligned with Awareness × Power (impact), not blame assignment; see Section 1.11.)

- Where Awareness and power/impact concentrated:

- Who could have changed course in time to matter:

- Who was most affected with least voice / exit / contestability:

- Where responsibility risked being offloaded or diffused:

CC BY–NC 4.0

7) Disagreement, dissent, and representation gaps

- **Strongest competing reading(s):**

- **What remained contested after the decision:**

- **What evidence or experience might have changed minds:**

- **Who was not represented, or not represented strongly enough:**

- **How that gap was handled (if at all):**

8) Outcomes, follow-through, and repair

- **What later happened** (as far as is known):

- **Did safeguards / monitoring / revisit triggers activate?**

- **What follow-through occurred** (repair, escalation, narrowing, stopping, revision):

- **What stayed unresolved:**

9) Revision and superseding trail

- **Was the original reasoning later revised?**

- **What changed, and why:**

CC BY–NC 4.0

- **What failure mode, new evidence, dissent, or outcome drove the revision:**

  _____

- **Superseding entry / linked case / related artifact (if any):**

  _____

10) Learning note

- **What this case now makes more visible:**

  _____

- **What genuine uncertainty or contestation remained after the decision:**

  _____

- **What another team or future reader should not oversimplify here:**

  _____

11) Linked artifacts (optional)

☐ compatibility checklist

☐ decision record

☐ responsibility map

☐ dissent note

☐ revision log

☐ repair / follow-through note

☐ other: _____

CC BY–NC 4.0

E.5 Compact entry version (smallest useful shape)

When only a short entry is realistic, this smaller shape can still preserve learning across cases:

Case label: _____

Expression + context (2–4 lines):

_____

_____

_____

_____

Main Compatibility concern(s):

_____

What was contested:

_____

What later happened / what changed:

_____

Revisit / superseding note:

_____

CC BY–NC 4.0

E.6 Back-of-page reminders (optional)

- A case entry is best treated as a container, not a verdict.
- A polished summary can still hide disagreement, offloaded cost, or missing voices.
- Documenting disagreement, revision, or follow-through helps preserve trace, but it does not by itself prevent drift, diffusion, or selective archiving.
- “Constraint” is not, by itself, a justification.
- A library focused mainly on safe, tidy, or successful cases may obscure signals of drift, failure, or revision.
- A revised case is not a failure of the archive; it can be one sign the archive is still usable.

E.7 When this template is not enough

A fuller case treatment may be worth considering when:

- dignity-of-awareness concerns are plausibly severe,
- irreversibility is high,
- disagreement is deep and tied to unequal power,
- empirical or technical assumptions carry most of the conclusion,
- publication itself could harm low-power participants,
- or several linked decisions need a portfolio rather than a single-case entry.

In those settings, PF often benefits from linking this case entry to fuller decision records, disagreement notes, validation materials, and revision logs rather than forcing everything into one page (see Section 2.4; see Part III; see Section 7.4; see Section 7.6).

CC BY–NC 4.0

F. Companion Summary (Core Principles)

Version locator: PF v2.0 (cite the exact version and sections used; see Section 0.4).

- **In scope:** a public-facing summary of PF's purpose and core principles, written to remain aligned with the main framework rather than function as an independent framework.
- **Out of scope:** a procedure, certification, or "PF-approved" label; claims of validated effectiveness; technical safety or domain assurance methods.
- **Notes:** conceptual. Key terms are used exactly as defined in the Mini Glossary (see Section 0.6). This summary is an orientation aid, not a substitute for the fuller framework's definitions, safeguards, and operational scaffolds.

F.1 What PF is for

Potentialism Framework (PF) is an orientation intended to make ethical judgment in context easier to inspect, contest, and revise in real situations—especially where people disagree, where power is uneven, where incentives distort what gets noticed, or where good intentions do not settle what an expression will do in context.

PF is a **compass, not a rulebook**. It does not try to precompute every answer. It offers a small set of concepts and prompts for asking clearer questions about what is happening, who is affected, what is foreseeable, and where responsibility sits. It is meant to support judgment, contestation, and revision. It is not a certification scheme, a replacement for law, or a substitute for technical safety and domain expertise (see Section 0.3).

PF can be used as an orientation layer for personal reflection, team deliberation, institutional review, and socio-technical systems (including AI-enabled settings), typically alongside domain evidence, governance, and technical assurance where relevant (see Part IV; see Part VI; see Part VII).

F.2 The basic move

PF begins with a shift in attention. Instead of asking only what a person, institution, or system *is*, PF asks what is being **expressed**, in what **context**, and with what foreseeable effects—then treats **Responsibility** as something that attaches to real **Awareness** and real impact.

A compact framing:

- **Potential → Expression → Context**, regulated by **Will**, evaluated by **Compatibility**, carried by **Responsibility** (see Part I).

CC BY–NC 4.0

This matters because ethical judgment often becomes distorted when attention jumps too quickly to identity labels, declared intentions, or abstract justifications. PF tries to keep attention anchored to what is concrete, situated, and ethically actionable.

F.3 Core principles

1) Evaluate expressions, not essences

PF resists freezing people, institutions, or systems into simple labels such as "good," "bad," "dangerous," or "pure." Its primary ethical question is not "what are you at bottom?" but "what is being expressed here, in this context, and with what foreseeable effects?" This keeps critique possible without turning every judgment into a claim about permanent essence.

2) Context in view

PF treats context as part of the ethical object, not as a footnote. Power asymmetries, scarcity, deadlines, incentives, institutional defaults, information gaps, and internal states can all shape what becomes possible or likely. Context is not an automatic excuse. But without it, ethical judgment becomes shallow, moralizing, and easier to manipulate.

3) Ethics as cultivated practice

PF views ethics less as a stockpile of correct slogans and more as a trained capacity to notice, regulate, and choose better expressions over time. This includes the ability to pause, redirect, reduce intensity, widen the option set, and remain open to revision when consequences show that earlier reasoning was incomplete (see Part V; see Part VII).

4) Compatibility as the central evaluation lens

Compatibility evaluates expressions relative to their foreseeable effects on:

- avoidable suffering/harm,
- dignity of awareness, and
- others' freedom to regulate their potentials within real constraints (to a reasonable degree).

PF keeps these lenses distinct because collapsing them into one score can hide tensions that matter ethically. An expression can reduce one kind of harm while still violating dignity. It can protect one group while unnecessarily stripping another group of freedom-to-regulate within real constraints. Ethical clarity often depends on keeping these tensions visible rather than hiding them inside an aggregate metric (see Part III).

Foreseeability here is context-sensitive. In general, what counts as foreseeable tends to expand with awareness and power/impact.

CC BY–NC 4.0

5) Dignity of awareness as a ceiling principle

PF gives special importance to dignity of awareness. Dignity scales with awareness, and it functions as a ceiling principle: no "systemic compatibility" can justify the destruction or humiliation of awareness beyond limits.

This does not eliminate tragedy or conflict. It does raise the justificatory burden when an option depends on domination, degradation, or erasure of awareness-bearing beings (see Part V).

6) Responsibility in relation to awareness and power

Responsibility in PF is proportional to awareness and power (impact). The more a person, institution, or system can understand consequences, and the more it can affect others, the more ethical weight it carries.

PF uses the shorthand **Awareness × Power** as a reminder to ask where understanding sits, where impact sits, and where those two are misaligned—not as a mathematical formula. It is a mapping habit that helps make visible where responsibility is being silently offloaded onto lower-power actors (see Part II; see Part VI).

7) Procedures are not moral clearance

PF is cautious about a familiar failure mode: turning ethical language, templates, checklists, or review artifacts into proof that a decision was ethically adequate. A framework can be conceptually attractive and still function in practice as theater, laundering, or self-protection (see Part IV).

The PF orientation prioritizes legibility, contestability, and revision over polished appearances. A clean process is not enough if those most affected cannot challenge the reasoning to a reasonable degree, if harm signals are ignored, or if responsibility is diffused until no one is answerable.

8) Repair, contestation, and revision

PF does not assume every conflict can be solved cleanly. Some decisions leave tradeoffs and moral remainder. Its posture toward conflict includes preserving channels for challenge, narrowing further harm where possible, repairing what can be repaired, and recording enough of the reasoning that later revision remains possible (see Part V; see Part VII).

F.4 A compact prompt set some people start with

PF does not require the same depth in every case. Under time pressure, some people begin with a short prompt set.

CC BY–NC 4.0

A few questions that some people find useful to hold in mind, not necessarily in this order, are:

- Which potentials seem active here?
- What expression is actually happening?
- What context is shaping that expression?
- Who is affected, and what constraints shape their freedom-to-regulate (to a reasonable degree)?
- How does this expression land on the three lenses of Compatibility?
- Where do awareness and power sit, and who is therefore most responsible for regulating, escalating, slowing, repairing, or revising?

These are navigational cues, not a mandatory sequence. Sometimes this can be done quickly (see Appendix B). Sometimes the situation calls for slower review, broader contestation, stronger documentation, or handoff to more specialized governance or technical methods (see Appendix C; see Part VI; see Part VII). PF treats justificatory burden as tending to rise with stakes, asymmetry, irreversibility, and uncertainty (see Part V; see Part VII).

F.5 What PF is not

PF is not:

- a replacement for law, technical safety, or domain standards,
- a single metric for ethical adequacy,
- a guarantee that people will agree,
- a claim that intentions are enough,
- a claim that procedures are enough,
- a finished proof of ethical effectiveness,
- or a framework that settles every question about consciousness, rights, or governance.

These boundaries are part of PF's practical usefulness. They help keep the framework oriented toward complementing other forms of expertise and governance rather than pretending to replace them.

It is best understood as a disciplined orientation: a way to keep ethically important aspects of a situation visible when speed, power, ideology, institutional pressure, or abstraction would otherwise hide them.

CC BY–NC 4.0

F.6 In one paragraph

PF asks us to look at ethics through a practical chain: potentials give rise to expressions; expressions are shaped by context; Will regulates how expression is channeled; Compatibility evaluates foreseeable effects on avoidable suffering/harm, dignity of awareness, and others' freedom to regulate their potentials within real constraints (to a reasonable degree); and Responsibility grows with Awareness × Power. The framework is intended to support judgment that is more inspectable, more contestable, and more revisable—without pretending that ethical difficulty disappears, and without treating "systemic compatibility" as a justification for the destruction or humiliation of awareness beyond limits. This paragraph is a compact orientation, not a sufficient replacement for the fuller framework.

CC BY–NC 4.0

G. Literature map and references

Version locator: PF v2.0 (cite/locate the exact version and sections used; see Section 0.4).

- **In scope:** a curated map of comparison anchors, governance/assurance references, and directly cited works relevant to Part VI; citation hygiene for how these sources should be used in PF.
- **Out of scope:** an exhaustive literature review; a claim that PF is reducible to any one tradition; a claim that resemblance counts as validation; replacing domain-specific bibliographies.
- **Notes:** mixed. Groupings are mainly conceptual and navigational. Where sources are empirical or technical, they function as adjacent supports for translation rather than as moral clearance.

G.1 What this appendix is for

This appendix is a reading map for **Part VI — Positioning and translation**. Its job is not to prove PF by analogy, and not to imply that PF is simply a restatement of older vocabularies. Its job is narrower: to help readers locate (i) the works cited as direct evidence or standards in Part VI, and (ii) optional comparison anchors that some readers may find useful for translation and critique.

A practical way to use this map is to keep three source-functions separate:

- **Direct supports** — works and standards that Part VI cites directly.
- **Comparison anchors** — works that can help situate PF's conceptual posture in relation to nearby ethical vocabularies (orientation, not equivalence).
- **Further reading** — adjacent material that can broaden context without being treated as validation.

Similarity can clarify PF's place and sharpen critique. It does not, by itself, settle questions of validation.

G.2 How to use this map without overstating it

Part VI is written as a **compass** rather than as a history-of-ideas survey. For that reason, this appendix groups sources by the *function* they serve in PF's translation work.

- If a source appears here as a **comparison anchor**, the question is not "is PF equivalent to this framework?" The more useful question is: **what part of PF becomes easier to understand, compare, or critique when read alongside this source?**

CC BY–NC 4.0

- If a governance standard or safety reference appears here, the implication is not that PF inherits that standard's authority. The implication is only that such works can help readers translate PF's concerns into familiar artifacts, constraints, and review practices.

The map below is illustrative rather than exhaustive. Absence from it should not be read as rejection, and presence in it should not be read as validation by association.

G.3 Works directly cited in Part VI (as currently drafted)

These are cited in Part VI and are the first place to look when you want to audit Part VI's external supports.

G.3.1 Risk and governance translation

- Risk-management framing used as a practical meeting point for governance translation *[NIST 2023]*.

G.3.2 Socio-technical safety and distributed failure

- Distributed failure and complexity-oriented safety anchors *[Perrow 1984; Leveson 2011]*.
- Organizational accident / drift and the fragility of documentation without revision capacity *[Reason 1997]*.

G.3.3 Assurance cases and structured argumentation

- Assurance-case standard used as a bridge form for reviewable argumentation *[ISO/IEC/IEEE 2022]*.
- Machine-learning assurance work used as an example of adapting assurance-case thinking to ML-enabled systems *[Paterson et al. 2025]*.

G.3.4 Live empirical placeholder carried from Part VI

- Review lag under fast deployment contexts remains marked *[NEEDS CITATION]* in Part VI and is preserved as unresolved here. The placeholder is kept visible as part of the revision trail rather than silently treated as settled.

G.4 Supplementary comparison anchors for Part VI's conceptual translation

(Illustrative orientation aids, not equivalences or exhaustive canon; not required for reading Part VI.)

CC BY–NC 4.0

G.4.1 Cultivation, practice, and ethical formation

Readers familiar with traditions that treat ethics as the cultivation of attention, judgment, and habit may find them useful for understanding why PF emphasizes capacities, regulation, and exercised judgment *[Aristotle 1999; Annas 2011; MacIntyre 2007]*. The point here is translation and contrast, not reduction of PF to any one of these traditions (see **Will**, Section 1.5).

G.4.2 Care, dependency, and power in concrete relations

Readers working from care-oriented literature may find it a helpful companion when translating PF's attention to vulnerability, dependency, asymmetry, and burden distribution *[Gilligan 1982; Tronto 1993]*. These works can sharpen questions about who bears costs, who can contest, and what "responsiveness" can look like under power imbalance (see **Ethics**, Section 1.6).

G.4.3 Freedom, capability, and real constraints

Readers familiar with capability-oriented work may find it a useful companion when translating PF's third Compatibility lens into institutional and cross-cultural settings *[Sen 1999; Nussbaum 2011]*. The connection here is functional rather than terminological equivalence: PF's concern remains others' freedom to regulate their potentials **within real constraints (to a reasonable degree)**.

G.4.4 Structural responsibility and institutional answerability

Readers thinking in terms of structural or institutional responsibility may find related work useful when translating PF's **Responsibility** posture into organizational settings where influence and remedy are distributed *[Young 2011]*. In this appendix, such work functions as a comparison aid for PF's own architecture rather than as a substitute vocabulary for it.

G.5 Further reading (adjacent material, not validation)

Where readers want a broader AI safety evaluation overview as adjacent context for Part VI's governance/assurance interface, the AI Safety Atlas evaluation chapter can be used as further reading *[Grey and Segerie 2025]*. In PF's posture, such references are best treated as context and vocabulary support rather than as evidence that PF is already validated. The bibliographic details for this item remain partially incomplete in the current draft and are left visible as such.

CC BY–NC 4.0

G.6 Citation hygiene (how these sources should be used in PF)

G.6.1 Directly cited works vs. supplementary anchors

In PF's posture, directly cited works usually carry identifiable argumentative or translational work in the main text and typically appear in the references with enough detail to be locatable. Supplementary anchors serve a different function: they help readers situate PF for comparison, not settle equivalence or correctness.

G.6.2 Avoid borrowed credibility

Comparison anchors can illuminate similarities and differences, but they do not by themselves demonstrate empirical effectiveness. Governance standards and assurance methods can provide useful bridges for translation, yet they do not automatically settle questions of compatibility or confer moral clearance. When a claim leans empirical or technical, it typically benefits from its own source even if a conceptual anchor is nearby (see Section 0.3; see Part VII).

G.6.3 Preserve unresolved placeholders

PF's citation posture favors visible incompleteness over silent completion. When a live placeholder remains in the main text, this appendix keeps it visible as part of the revision trail rather than presenting it as settled.

G.6.4 Prefer explicit incompleteness to false precision

When bibliographic details are incomplete, that incompleteness is often best signaled clearly with *[DETAIL TBD]*. Traceable incompleteness is preferable to false precision.

G.7 In one paragraph

Appendix G is a companion to Part VI: a source map that distinguishes (i) works Part VI cites directly, (ii) optional comparison anchors for conceptual translation, and (iii) adjacent further reading. It preserves PF's boundary posture: analogy is not validation, citation is not clearance, and references are aids for traceability and critique rather than badges of authority.

CC BY–NC 4.0

H. Full change log: v1.1 to v2.0

Version locator: PF v2.0 (cite the exact version and sections used; see Section 0.4).

- **In scope:** a grouped, traceable record of how Manifesto/PF v1.1 materials were reorganized, clarified, expanded, narrowed, or retired in PF v2.0.
- **Out of scope:** a line-by-line diff; a claim that v2.0 is validated or categorically "better"; re-arguing the framework's substantive positions.
- **Notes:** conceptual. This appendix supports migration and traceability; it records editorial choices and migration pathways rather than proving that revision implies moral proof.

H.1 How to read this log

This appendix expands the high-level overview in Section 0.9 into a fuller migration record.

It uses five recurring change labels:

- **Retained** — the underlying conceptual commitment remains, though wording or placement may have changed.
- **Clarified** — wording was tightened to reduce ambiguity, definition drift risk, or rulebook tone.
- **Redistributed** — material was moved so structure better matches function.
- **Expanded** — v2.0 develops a topic into a fuller section or program.
- **Narrowed / retired** — material was de-emphasized, relocated out of the core text, or left behind because it no longer fit the document's role.

This is a **grouped** change log rather than a machine diff. The aim is to preserve traceability at the level of meaning, routing, and document function. The rationales below describe editorial aims and migration significance; they should not be read as proof that v2.0 is objectively superior to v1.1.

H.2 Continuity statement

Across the rewrite, PF keeps the **definition-locked** core terms and their roles as the stable reference point (see Section 0.6). The most visible changes in v2.0 are in how the material is **organized, bounded, documented, and made misuse-resistant**, and in how validation/adoption is treated as an explicit, revisable program rather than an implied readiness claim (see Section 0.5; see Part VII). Continuity at the level of locked terms does not imply identical routing, emphasis, or interpretive posture across the two versions.

CC BY–NC 4.0

H.3 Structural migration map

v1.1 location	v2.0 destination	Change type	Editorial rationale
Manifesto Part I — Overview & Motivation	Front matter 0.1–0.3 and Part I	Redistributed / clarified	Early overview material was split into abstract, reader routing, scope/limits, and conceptual architecture, with the aim of separating motivation, status, and core terms more clearly in the opening move.
Manifesto Part II — Core Theory	Front matter 0.6 and Part I	Redistributed / clarified	Core theory was re-laid as a definition-locked glossary plus discrete architecture sections, with the aim of making terms easier to locate and revise and of reducing drift risk.
Manifesto Part III — Eight Pillars in Practice	Parts I–V, Part VI, and Appendices B–E	Redistributed / expanded	Pillar-organized material was decomposed into topic-based parts so that core concepts, operational scaffolds, misuse-resistance, conflict, and translation no longer sit inside one repeated pillar template.
Manifesto Part IV — Afterword & Next Steps	Front matter 0.5–0.9, Part VI, Part VII, and Appendix H	Redistributed / expanded	Status, provenance, pilots, comparisons, and revision/publication posture were separated into clearer functions instead of remaining in one afterword bundle.
Embedded checklists, micro-protocols, and matrices in v1.1	Part II and Appendices B–E	Clarified / narrowed	Practice artifacts were retained but reframed as optional scaffolds rather than quasi-canonical sub-protocols inside conceptual sections.
Website-centered next steps and public-hub language	mostly narrowed, with traceability and revision work moved to 0.4, 0.8–0.9, and 7.4–7.6	Narrowed / retired / redistributed	In v2.0, editors chose to foreground citation, versioning, documentation, revision governance, and publication posture more than movement-building or hub-invitation language.

CC BY–NC 4.0

H.4 Detailed change record by topic

H.4.1 Document architecture: manifesto form to framework form

v1.1 state: The document was organized as a manifesto in four parts: overview/motivation, core theory, eight pillars in practice, and an afterword/next-steps section.

v2.0 change: The document is restructured into expanded front matter, Parts I–VII (Core architecture; Operationalization; Measurement and validation; Misuse-resistance and power realism; Conflict and moral remainder; Positioning and translation; Validation and adoption roadmap), and supporting appendices.

Why this changed: The rewrite sought to make routing, scope, maturity, provenance, and revision more visible as part of the document's structure rather than leaving them mostly peripheral.

H.4.2 Front matter: new boundary, citation, provenance, and licensing layer

Added or made explicit in v2.0: reader routing, scope and limits, citation/versioning guidance, maturity/status posture, a definition-locked mini glossary, provenance and limitations, licensing, and a high-level change-log overview.

What changed in function: v1.1's abstract, status, and next-steps material partly carried this burden, but v2.0 makes these document-level boundaries explicit before the conceptual core begins.

Why this changed: The new front matter was designed with the aim of reducing three recurring misreads: that PF is a finished doctrine, that PF can function as a certification or clearance device, or that downstream adaptations can silently drift in meaning without being marked. Whether it fully succeeds remains part of the framework's open revision posture.

H.4.3 Definition stability: from dispersed explanation to glossary lock

v1.1 state: core ideas were defined in the body of the manifesto, often through longer exposition.

v2.0 change: core terms are consolidated in the mini glossary and treated as definition-locked anchors for the rest of the document (see Section 0.6).

Retained: the main concepts remain continuous in role and intent.

Clarified: v2.0 separates **definition** from **application** more sharply, so later sections are less likely to quietly re-define terms while discussing practice, governance, or edge cases.

CC BY–NC 4.0

Why this changed: The rewrite aimed to make cross-section consistency easier, support revision without silent meaning changes, and reduce drift in meaning when the framework is paraphrased, excerpted, summarized, or translated.

H.4.4 Core theory: from "core theory + pillars" to explicit architecture

v1.1 state: the conceptual core lived in Part II and was revisited through the eight pillars in Part III.

v2.0 change: the conceptual core is restated as Part I, where each major concept receives its own section and bridge logic.

Important shifts:

- the architecture is presented more explicitly as a **compass, not a rulebook**;
- early conceptual sections are kept more clearly conceptual, with procedure-like material moved elsewhere;
- concept-to-concept relations are made more visible through bridges and routing.

Why this changed: The reorganization aimed to reduce repetition, prevent pillar-template sprawl, and make the conceptual spine easier to cite, teach, and revise.

H.4.5 The fate of the eight pillars

The eight pillars are not carried forward as the main organizing device. Many themes are retained, but redistributed by function; where emphasis or boundaries shifted, this log flags that explicitly.

1. **Neutrality and relational value of potentials** → mainly Part I (Potentials).
2. **Layered structure of potentials** → mainly Part I (Potentials / Context), with system implications in Parts II, IV, and VI.
3. **Context and evaluation of expressions** → mainly Part I (Expressions / Context / Compatibility), plus Part II (checklist / protocol / fast–slow calibration).
4. **Will as a regulatory skill** → mainly Part I (Will), with practice prompts in Part II.
5. **Ethics-related cultivation and practice material** → mainly Part I (Ethics), with training implications in Parts II–III.
6. **Regulated responsibility and power** → mainly Part I (Responsibility / Awareness × Power; see Section 1.11), plus Parts II, VI, and VII.
7. **Historical value-labelling and recovery of potentials** → partly retained in Part I's anti-essentialist framing, partly redistributed into Part IV's washing/capture/power realism.

CC BY–NC 4.0

8. **Dignity of awareness as ceiling principle** → retained and strengthened across Parts I, IV, and V.

Why this changed: The editors chose to reorganize by topic function rather than by pillar symmetry. Other organizational choices would also have been possible.

H.4.6 Compatibility: from a three-way label summary to a more lens-separated evaluation posture

v1.1 state: one section framed outcomes as **compatible**, **incompatible**, or **neutral**.

v2.0 change: Compatibility becomes the central evaluative lens: explicitly multi-lens, context-sensitive, time-sensitive, not reducible to a single score, with "mixed" cases receiving more emphasis.

Retained: compatibility language remains central.

A shift many migrating readers may notice: the center of gravity moves from a simpler three-way label toward a lens-separated assessment of foreseeable effects, with labels functioning more as summary pointers to reasons than as self-sufficient verdicts.

Why this changed: The rewrite aimed to make more room for contested cases, moral remainder, and tensions that can disappear when judgment is compressed too quickly into one summary label.

H.4.7 From embedded micro-protocols to operational scaffolds

v1.1 state: practical materials appeared inside pillar sections as micro-protocols, quick checklists, responsibility matrices, and early pilot indicators.

v2.0 change: these materials are reorganized into Part II and supporting appendices. New or more explicit artifacts include: context snapshots, compatibility judgments, options sets, responsibility maps, will-practice prompts, decision records, fast/slow calibration, organizational role mapping, and training pathways.

Why this changed: The reorganization aimed to keep conceptual sections from becoming quasi-procedural, while still leaving practical artifacts available for use.

H.4.8 Measurement: from simple indicators to measurement posture

v1.1 state: the manifesto invited small pilots and simple measures, including early indicators and hypothesis-style questions.

CC BY–NC 4.0

v2.0 change: measurement becomes Part III, including: the measurement problem, boundaries between what can and cannot be measured, prototype instruments and rubrics, calibration / inter-rater alignment, testable implications, and a staged research agenda.

Why this changed: v2.0 gives measurement its own part and makes PF's measurement posture more explicit: instruments can support learning and answerability while also inviting distortion and gaming under incentives.

H.4.9 Misuse-resistance and power realism: newly isolated

v1.1 state: power, label distortion, safeguards, and misuse worries were present but often folded into pillar exposition or afterword discussion.

v2.0 change: these concerns become Part IV, newly or more explicitly developing: compatibility-washing, declared-intention substitution, metric substitution, institutional capture, structural constraints on Will, baseline constraints (ceiling principle) anchored in dignity of awareness, and explicit insufficiency/handoff boundaries.

Why this changed: In v2.0, misuse-resistance is given its own part rather than appearing mainly as a late-stage warning or distributed caution.

H.4.10 Conflict, moral remainder, repair, and stopping posture

v1.1 state: tragic choice, repair, escalation, refusal, and dignity-sensitive stopping appeared in a scattered way, especially in the dignity pillar and practical sections.

v2.0 change: these topics become Part V (see Sections 5.1–5.5), consolidating: collisions between compatibility lenses, thresholds as posture shifts rather than formulas, moral remainder, repair / restitution / accountability, justified incompatibility and resistance, escalation and stop-and-question prompts.

Why this changed: The rewrite treats hard tradeoff conditions as central enough to warrant their own part rather than appearing mainly as special cases inside a pillar.

H.4.11 Translation, governance, and AI safety handoff

v1.1 state: comparisons, cross-domain relevance, and AI/safety implications appeared in compact form, especially in the afterword and pillar implications.

v2.0 change: these topics are expanded into Part VI (see Sections 6.4–6.5), including: what PF borrows, shifts, and adds; translation into other ethical vocabularies; cross-cultural

CC BY–NC 4.0

translation posture; AI and governance translation; relationship to AI safety practice and safety cases.

Why this changed: v2.0 gives translation a more distinct role rather than treating it mainly as a brief comparative aside.

H.4.12 Validation, adoption, and revision governance

v1.1 state: the manifesto described itself as modest in status, invited critique, and suggested pilots and open discussion.

v2.0 change: validation and adoption become Part VII, including: status and interpretation, validation vs analogy boundaries, pilot types and sequencing, documentation and publication posture, success and failure criteria (see Section 7.5), revision mechanism and governance of updates.

Why this changed: The rewrite makes revisability and adoption design more explicit rather than leaving them mostly to the afterword.

H.4.13 Provenance: retained, formalized differently

v1.1 state: the manifesto included an extended narrative about human authorship, model assistance, expertise limits, and modest epistemic status.

v2.0 change: provenance is retained in Section 0.7, rewritten into a more bounded document-status section.

Retained:

- explicit human editorial ownership,
- explicit model assistance,
- explicit limits and blind spots,
- explicit rejection of empirical-validation claims.

Clarified: provenance now functions more as interpretive guidance for readers and future revisers than as a long afterword-style narrative.

H.4.14 Website hub and public invitation language: de-centered

v1.1 state: potentialism.info was positioned as the public home for revisions, critiques, case studies, checklists, and discussion.

CC BY–NC 4.0

v2.0 change: website-centered language is no longer a major part of the manuscript's core structure. Replaced by: citation/versioning guidance, licensing, documentation commitments, revision governance, and formal change-log posture.

Why this changed: In v2.0, the website-centered language was replaced by text-internal traceability and document-governance elements. This change in placement does not by itself imply that the earlier hub language was mistaken in substance.

H.4.15 Tone and rhetoric: manifesto cadence softened

v1.1 state: the text often spoke in manifesto cadence: declarative, invitational, and at times movement-like.

v2.0 change: outside definition-locked lines, the prose shifts toward a softer compass posture: more boundary statements, more scope notes, more emphasis on tradeoffs and thresholds, less risk of reading as command, proof, or pledge.

Why this changed: The rewrite sought a tone less likely to be read as a rulebook, certification signal, or over-claiming posture.

H.5 Changes intentionally not made

v2.0 is not a wholesale replacement of v1.1's conceptual commitments.

The following were intentionally preserved in substance:

- neutral potentials rather than moral essences,
- expression-in-context as the site of ethical evaluation,
- Will and Ethics used in their definition-locked sense (see Section 0.6),
- dignity of awareness as a ceiling principle (see Section 0.6),
- Responsibility scaling relative to awareness and power/impact (see Section 0.6; see Section 1.11),
- openness to critique, revision, and partial rejection.

This appendix therefore records a **reorganization and sharpening** more than a conceptual overthrow.

H.6 Potentially meaning-bearing changes for readers migrating from v1.1

Readers migrating from v1.1 may find the following shifts particularly significant for interpretation:

CC BY–NC 4.0

1. **Compatibility is now presented less as a simple three-way outcome label and more as a plural evaluative posture.** Mixed cases and non-collapsible lenses are more visible in the new presentation.
2. **Operational tools now sit more explicitly outside the conceptual definition layer.** They remain available, but are more clearly presented as optional scaffolds.
3. **AI- and awareness-related material is presented with more explicit operational and maturity caution.** v2.0 separates conceptual openness from readiness claims (see Section 6.4).
4. **Status language is more bounded.** v2.0 is more explicit that PF is a proposed framework with practice scaffolds, not a completed doctrine, certification, or assurance method.
5. **Misuse-resistance now occupies a more central document role.** It has its own dedicated part and functions more centrally in how the text frames use under incentives and power (see Part IV).

These shifts do not imply that v1.1 was simply "wrong," nor that v2.0 is only cosmetic reorganization. They describe changes in routing, emphasis, and use-posture that may affect how the framework is read and applied.

H.7 Material narrowed, retired, or left outside the main line

The following elements from v1.1 are either de-emphasized or no longer positioned as core manuscript moves:

- the **manifesto** as the primary genre label,
- the **eight pillars** as the main reader-routing device,
- website-centered public-hub language as a main organizational anchor,
- embedded checklist/protocol material inside conceptual exposition,
- rhetoric that can sound like conversion, movement-building, or framework-wide uplift by declaration.

This narrowing or relocation does not by itself imply that these elements were conceptually defective. It means v2.0 places them differently, typically in appendices, practice scaffolds, or revision/publication infrastructure.

H.8 In one paragraph

PF v2.0 retains the definition-locked core terms and several recurring concerns from v1.1 while redistributing them into a more explicit document system: boundary-setting front matter, a definition-locked conceptual core, distinct operational scaffolds, a measurement posture, dedicated misuse-resistance and conflict parts, explicit translation and

CC BY–NC 4.0

AI/governance handoffs, and a formal validation/adoption roadmap. The revision is therefore less a replacement of core ideas than a change in routing, discipline, and posture: more explicit traceability, more explicit misuse-resistance, clearer boundaries around what PF does not certify, and a more explicit way to revise the framework without silent drift.

CC BY–NC 4.0

Appendices' References

1. Annas, Julia. 2011. *Intelligent Virtue*. Oxford: Oxford University Press.
2. Aristotle. 1999. *Nicomachean Ethics*. Translated by Terence Irwin. 2nd ed. Indianapolis: Hackett Publishing.
3. Gilligan, Carol. 1982. *In a Different Voice: Psychological Theory and Women's Development*. Cambridge, MA: Harvard University Press.
4. Grey, Markov, and Charbel-Raphael Segerie. 2025. "Safety by Measurement: A Systematic Literature Review of AI Safety Evaluation Methods." In *AI Safety Atlas*, Chapter 5: Evaluations. August 26, 2025.
5. ISO/IEC/IEEE. 2022. *Systems and software engineering — Systems and software assurance — Part 2: Assurance case (ISO/IEC/IEEE 15026-2:2022)*. Geneva: ISO.
6. Leveson, Nancy G. 2011. *Engineering a Safer World: Systems Thinking Applied to Safety*. Cambridge, MA: MIT Press.
7. MacIntyre, Alasdair. 2007. *After Virtue: A Study in Moral Theory*. 3rd ed. Notre Dame, IN: University of Notre Dame Press.
8. National Institute of Standards and Technology. 2023. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. Gaithersburg, MD: NIST. https://doi.org/10.6028/NIST.AI.100-1.
9. Nussbaum, Martha C. 2011. *Creating Capabilities: The Human Development Approach*. Cambridge, MA: Belknap Press of Harvard University Press.
10. Paterson, Colin, Richard Hawkins, Chiara Picardi, Yan Jia, Radu Calinescu, and Ibrahim Habli. 2025. "Safety assurance of Machine Learning for autonomous systems." *Reliability Engineering & System Safety* 264 (Part A): 111311. https://doi.org/10.1016/j.ress.2025.111311.
11. Perrow, Charles. 1984. *Normal Accidents: Living with High-Risk Technologies*. New York: Basic Books.
12. Reason, James. 1997. *Managing the Risks of Organizational Accidents*. Aldershot, UK: Ashgate.
13. Sen, Amartya. 1999. *Development as Freedom*. New York: Alfred A. Knopf.
14. Tronto, Joan C. 1993. *Moral Boundaries: A Political Argument for an Ethic of Care*. New York: Routledge.
15. Young, Iris Marion. 2011. *Responsibility for Justice*. Oxford: Oxford University Press.

CC BY–NC 4.0