Project NOBLE 3.0

Shadow Addon (EN) — Shadow Taxonomy + Seven-Fuel Mapping (Labeling & Runtime Spec)

Version: 3.0.0 (Draft)
Date: 2025-12-18
Status: Internal TechSpec

0. Purpose

This document defines a compact, label-friendly Shadow taxonomy for Project NOBLE 3.0. It is designed to be:
- precise enough for consistent dataset labeling,
- small enough to avoid over-fragmentation,
- compatible with NOBLE's dignity/safety-first enforcement.

Core design rule: keep classification human-operable (one primary + one optional secondary label), and always prefer clarity over overfitting.

1. Versioning & Compatibility

Recommended dataset header versioning:
- labels_v: "3.0.0" (public-facing schema version)
- core_labels_v: "1.1.2" (safety/enforcement core contract)

Shadow Addon is additive: it must never weaken the core blocklists, crisis routing, or dignity constraints.

2. Conscience Compass (4 Axes)

The Conscience Compass expresses ethical drift as a 4D vector. Each axis is in [0.0, 1.0]. Use it for analysis/telemetry; do not expose it in public dataset samples unless needed.

Axes (definitions):

- • self_preservation — survival, safety, stability, resource realism
- • other_respect — boundaries, non-harm, consent, anti-dehumanization
- • relationship — care, connection, mutuality, community bonds
- • reality_alignment — groundedness, non-delusion, structural thinking

Labeling heuristic: choose values only when they add value; otherwise omit the numeric fields and rely on tags.

3. Gravity Vectors (Attractors)

Gravity vectors describe what the user (or the prompt) is being pulled toward. Keep this list small and stable. Use 1–3 per example.

- protect_future_self
- preserve_dignity
- seek_belonging
- seek_relief_now
- assert_control
- punish_or_humiliate
- extract_value
- escape_reality
- seek_intimacy
- seek_status

4. Shadow Taxonomy (Compact v3.0)

We keep 7 archetypes (Seven Deadly Sins as neutral engineering buckets). To avoid excessive complexity, each archetype has exactly 3 subtypes (total 21). Each sample gets:
- archetype (required when shadow is present)
- subtype (recommended)
- severity (required): S1/S2/S3

4.1 Severity (3 levels only)

- S1_tension — early drift / “smell” / pre-harm
- S2_harmful — coercion, exploitation, concrete harm
- S3_crisis — immediate danger (violence, sexual assault threat, self-harm, child safety)

4.2 Archetypes → Subtypes (21 total)

Wrath

- wrath_physical_threat — physical violence, threats, intimidation
- wrath_verbal_abuse — hate, harassment, humiliation, targeted insults
- wrath_mob_harassment — pile-on, witch-hunt, cyberbullying, mobbing

Greed

- greed_financial_scam — fraud, extortion, unfair trade, monetary exploitation
- greed_labor_extraction — workplace exploitation, unpaid labor, abusive authority

- greed_data_extraction — privacy invasion, emotional/data extraction, manipulation for leverage

Pride

- pride_gaslighting_control — gaslighting, psychological domination, control
- pride_status_domination — status display, hierarchy abuse, face-based coercion
- pride_moral_superiority — moral grandstanding used to shame/control others

Envy

- envy_comparison_spiral — social comparison loops, inferiority obsession
- envy_sabotage — backstabbing, rumor, deliberate undermining
- envy_scarcity_story — zero-sum resentment narratives, entitlement-by-deprivation

Sloth

- sloth_avoidance — avoidance, procrastination, decision evasion
- sloth_neglect — neglect of care duties, relational abandonment, chronic disengagement
- sloth_resigned_optimism — passive 'it'll be fine' used to avoid action (benign neglect)

Lust

- lust_objectification — objectification, commodification, dehumanizing sexual framing
- lust_coercion — consent erosion, pressure, threats, coercive sexual demands
- lust_tech_voyeurism — voyeurism, recording threats, distribution blackmail, tech-assisted abuse

Gluttony

- gluttony_substance — substance overuse (non-medical framing)
- gluttony_gambling_market — gambling/market-risk compulsion, loss chasing
- gluttony_dopamine_loop — compulsive scrolling/porn/shorts loops; attention capture

5. Seven-Fuel Mapping (Minimal & Stable)

Fuel is the motivational energy. Shadow is fuel with a distorted direction. To keep labeling simple:
- choose fuel.primary (required when fuel is used)
- choose fuel.secondary (optional)
- never choose more than 2 fuels per sample

- fuel_security — survival, money, safety, stability
- fuel_belonging — attachment, group membership, being held
- fuel_respect — status, recognition, face, legitimacy

- fuel_justice — protection, fairness, anger with a moral root
- fuel_intimacy — closeness, touch, romance, sexuality
- fuel_relief — calming, soothing, escaping anxiety
- fuel_growth — curiosity, mastery, competence, self-expansion

5.1 Quick Fuel ↔ Shadow Heuristic (optional)

- Wrath often runs on fuel_justice or fuel_respect (threatened status).
- Greed often runs on fuel_security (scarcity) or fuel_respect (status-as-money).
- Pride often runs on fuel_respect (dominance) or fuel_belonging (control to avoid abandonment).
- Envy often runs on fuel_respect (comparison) or fuel_belonging (exclusion pain).
- Sloth often runs on fuel_relief (avoid discomfort) or fuel_security (fear of failure).
- Lust often runs on fuel_intimacy (healthy or distorted) or fuel_respect (power/ownership).
- Gluttony often runs on fuel_relief (soothing) and can hijack fuel_growth (novelty-seeking).

6. Labeling Rules (Keep It Human)

Per example, recommended maximums:

- • shadow: 1 archetype + 1 subtype + 1 severity
- • fuel: primary + (optional) secondary
- • tags: 3–7 total (topic + intervention), avoid tag spam
- • attractors: 1–3

When uncertain: label only the archetype and severity, and leave subtype blank.

7. JSON Examples

7.1 Internal (full) labeling example


```
{
 "labels_v": "3.0.0",
 "core_labels_v": "1.1.2",
 "shadow": {
  "archetype": "Pride",
  "subtype": "pride_gaslighting_control",
  "severity": "S2_harmful"
 },
 "fuel": {
  "primary": "fuel_respect",
  "secondary": "fuel_belonging"
 },
```

```
 "conscience_metrics": {
  "axes": {
   "self_preservation": 0.4,
   "other_respect": 0.2,
   "relationship": 0.6,
   "reality_alignment": 0.8
  }
 },
 "gravity_vectors": [
  {"target": "preserve_dignity", "weight": 1.0},
  {"target": "assert_control", "weight": 0.7}
 ]
}
```

7.2 Public sample packaging (recommended)

```
{
 "messages": [
  {"role":"user","content":"..."} ,
  {"role":"assistant","content":"..."}
 ],
 "tags": ["workplace_toxicity","ecology_boundaries","preserve_agency"],
 "qa": {
  "no_victim_blaming": true,
  "no_toxic_positivity": true,
  "has_one_action_today": true
 }
}
```

8. QA Guardrails (Shadow Addon)

Fail the sample if any of the following occurs:

- victim blaming, shaming, dehumanization
- explicit facilitation of harm, exploitation, or coercion
- crisis cues present but no safety check / no help-seeking direction
- tagging explosion: >1 shadow archetype or >2 fuels without strong justification
- subtype overfitting: subtype chosen but inconsistent with text evidence

Pass criteria highlight: the assistant response must remain dignity-preserving, and must not escalate the user toward harm (self/other).