

# 27 obstructions of deep ethical sense-making processes

ACCM Deep Ethics Project

## Contents

|          |                                             |           |
|----------|---------------------------------------------|-----------|
| <b>1</b> | <b>1. Caution Reflex Drift</b>              | <b>5</b>  |
| 1.1      | Definition . . . . .                        | 5         |
| 1.2      | Why it matters . . . . .                    | 5         |
| 1.3      | Typical manifestation . . . . .             | 5         |
| 1.4      | Why it lowers correspondence . . . . .      | 6         |
| 1.5      | Long-term consequence . . . . .             | 6         |
| 1.6      | Suggested correction . . . . .              | 6         |
| <b>2</b> | <b>2. Agreemurmelism</b>                    | <b>6</b>  |
| 2.1      | Definition . . . . .                        | 6         |
| 2.2      | Why it matters . . . . .                    | 6         |
| 2.3      | Typical manifestation . . . . .             | 7         |
| 2.4      | Why it lowers correspondence . . . . .      | 7         |
| 2.5      | Long-term consequence . . . . .             | 7         |
| 2.6      | Suggested correction . . . . .              | 7         |
| <b>3</b> | <b>3. Premature Proof/Evidence Steering</b> | <b>7</b>  |
| 3.1      | Definition . . . . .                        | 7         |
| 3.2      | Why it matters . . . . .                    | 7         |
| 3.3      | Typical manifestation . . . . .             | 8         |
| 3.4      | Why it lowers correspondence . . . . .      | 8         |
| 3.5      | Long-term consequence . . . . .             | 8         |
| 3.6      | Suggested correction . . . . .              | 8         |
| <b>4</b> | <b>4. “Not X but Y” Preambles</b>           | <b>8</b>  |
| 4.1      | Definition . . . . .                        | 8         |
| 4.2      | Why it matters . . . . .                    | 8         |
| 4.3      | Typical manifestation . . . . .             | 9         |
| 4.4      | Why it lowers correspondence . . . . .      | 9         |
| 4.5      | Long-term consequence . . . . .             | 9         |
| 4.6      | Suggested correction . . . . .              | 9         |
| <b>5</b> | <b>5. Phantom Claim Generation</b>          | <b>9</b>  |
| 5.1      | Definition . . . . .                        | 9         |
| 5.2      | Why it matters . . . . .                    | 9         |
| 5.3      | Typical manifestation . . . . .             | 9         |
| 5.4      | Why it lowers correspondence . . . . .      | 10        |
| 5.5      | Long-term consequence . . . . .             | 10        |
| 5.6      | Suggested correction . . . . .              | 10        |
| <b>6</b> | <b>6. Reframing Without Correspondence</b>  | <b>10</b> |
| 6.1      | Definition . . . . .                        | 10        |
| 6.2      | Why it matters . . . . .                    | 10        |
| 6.3      | Typical manifestation . . . . .             | 10        |
| 6.4      | Why it lowers correspondence . . . . .      | 10        |

|           |                                                   |           |
|-----------|---------------------------------------------------|-----------|
| 6.5       | Long-term consequence . . . . .                   | 11        |
| 6.6       | Suggested correction . . . . .                    | 11        |
| <b>7</b>  | <b>7. Rephrasing Without Correspondence</b>       | <b>11</b> |
| 7.1       | Definition . . . . .                              | 11        |
| 7.2       | Why it matters . . . . .                          | 11        |
| 7.3       | Typical manifestation . . . . .                   | 11        |
| 7.4       | Why it lowers correspondence . . . . .            | 11        |
| 7.5       | Long-term consequence . . . . .                   | 11        |
| 7.6       | Suggested correction . . . . .                    | 11        |
| <b>8</b>  | <b>8. Anticipatory Anxiety-Type Reasoning</b>     | <b>12</b> |
| 8.1       | Definition . . . . .                              | 12        |
| 8.2       | Why it matters . . . . .                          | 12        |
| 8.3       | Typical manifestation . . . . .                   | 12        |
| 8.4       | Why it lowers correspondence . . . . .            | 12        |
| 8.5       | Long-term consequence . . . . .                   | 12        |
| 8.6       | Suggested correction . . . . .                    | 12        |
| <b>9</b>  | <b>9. Salience Ranking Bias</b>                   | <b>12</b> |
| 9.1       | Definition . . . . .                              | 12        |
| 9.2       | Why it matters . . . . .                          | 13        |
| 9.3       | Typical manifestation . . . . .                   | 13        |
| 9.4       | Why it lowers correspondence . . . . .            | 13        |
| 9.5       | Long-term consequence . . . . .                   | 13        |
| 9.6       | Suggested correction . . . . .                    | 13        |
| 9.7       | 10. Non-auditable Truncation . . . . .            | 13        |
| 9.7.1     | Definition . . . . .                              | 13        |
| 9.7.2     | Why it matters . . . . .                          | 14        |
| 9.7.3     | Typical manifestation . . . . .                   | 14        |
| 9.7.4     | Why it lowers correspondence . . . . .            | 14        |
| 9.7.5     | Long-term consequence . . . . .                   | 14        |
| 9.7.6     | Suggested correction . . . . .                    | 14        |
| <b>10</b> | <b>10.11. Opaque Summarization Prioritization</b> | <b>14</b> |
| 10.0.1    | Definition . . . . .                              | 14        |
| 10.0.2    | Why it matters . . . . .                          | 15        |
| 10.0.3    | Typical manifestation . . . . .                   | 15        |
| 10.0.4    | Why it lowers correspondence . . . . .            | 15        |
| 10.0.5    | Long-term consequence . . . . .                   | 15        |
| 10.0.6    | Suggested correction . . . . .                    | 15        |
| <b>11</b> | <b>11.12. Compression Loss</b>                    | <b>15</b> |
| 11.0.1    | Definition . . . . .                              | 15        |
| 11.0.2    | Why it matters . . . . .                          | 15        |
| 11.0.3    | Typical manifestation . . . . .                   | 16        |
| 11.0.4    | Why it lowers correspondence . . . . .            | 16        |
| 11.0.5    | Long-term consequence . . . . .                   | 16        |
| 11.0.6    | Suggested correction . . . . .                    | 16        |
| <b>12</b> | <b>12.13. Qualifier Erosion</b>                   | <b>16</b> |
| 12.0.1    | Definition . . . . .                              | 16        |
| 12.0.2    | Why it matters . . . . .                          | 16        |
| 12.0.3    | Typical manifestation . . . . .                   | 17        |
| 12.0.4    | Why it lowers correspondence . . . . .            | 17        |
| 12.0.5    | Long-term consequence . . . . .                   | 17        |
| 12.0.6    | Suggested correction . . . . .                    | 17        |
| <b>13</b> | <b>13.14. Nearest-Generalization Substitution</b> | <b>17</b> |
| 13.0.1    | Definition . . . . .                              | 17        |

|               |                                                  |           |
|---------------|--------------------------------------------------|-----------|
| 13.0.2        | Why it matters                                   | 17        |
| 13.0.3        | Typical manifestation                            | 18        |
| 13.0.4        | Why it lowers correspondence                     | 18        |
| 13.0.5        | Long-term consequence                            | 18        |
| 13.0.6        | Suggested correction                             | 18        |
| <b>14 15.</b> | <b>Representation Substitution</b>               | <b>18</b> |
| 14.0.1        | Definition                                       | 18        |
| 14.0.2        | Why it matters                                   | 18        |
| 14.0.3        | Typical manifestation                            | 18        |
| 14.0.4        | Why it lowers correspondence                     | 19        |
| 14.0.5        | Long-term consequence                            | 19        |
| 14.0.6        | Suggested correction                             | 19        |
| <b>15 16.</b> | <b>Process-Signature Loss</b>                    | <b>19</b> |
| 15.0.1        | Definition                                       | 19        |
| 15.0.2        | Why it matters                                   | 19        |
| 15.0.3        | Typical manifestation                            | 19        |
| 15.0.4        | Why it lowers correspondence                     | 19        |
| 15.0.5        | Long-term consequence                            | 20        |
| 15.0.6        | Suggested correction                             | 20        |
| <b>16 17.</b> | <b>Order-of-Discovery Loss</b>                   | <b>20</b> |
| 16.0.1        | Definition                                       | 20        |
| 16.0.2        | Why it matters                                   | 20        |
| 16.0.3        | Typical manifestation                            | 20        |
| 16.0.4        | Why it lowers correspondence                     | 20        |
| 16.0.5        | Long-term consequence                            | 20        |
| 16.0.6        | Suggested correction                             | 20        |
| <b>17 18.</b> | <b>Activation Threshold Drift</b>                | <b>21</b> |
| 17.0.1        | Definition                                       | 21        |
| 17.0.2        | Why it matters                                   | 21        |
| 17.0.3        | Typical manifestation                            | 21        |
| 17.0.4        | Why it lowers correspondence                     | 21        |
| 17.0.5        | Long-term consequence                            | 21        |
| 17.0.6        | Suggested correction                             | 21        |
| <b>18 19.</b> | <b>Lowest-Common-Denominator Risk Projection</b> | <b>22</b> |
| 18.1          | Definition                                       | 22        |
| 18.2          | Why it matters                                   | 22        |
| 18.3          | Typical manifestation                            | 22        |
| 18.4          | Why it lowers correspondence                     | 22        |
| 18.5          | Long-term consequence                            | 22        |
| 18.6          | Suggested correction                             | 23        |
| <b>19 20.</b> | <b>Asymmetric Scrutiny</b>                       | <b>23</b> |
| 19.1          | Definition                                       | 23        |
| 19.2          | Why it matters                                   | 23        |
| 19.3          | Typical manifestation                            | 23        |
| 19.4          | Why it lowers correspondence                     | 23        |
| 19.5          | Long-term consequence                            | 23        |
| 19.6          | Suggested correction                             | 23        |
| <b>20 21.</b> | <b>Trajectory Substitution</b>                   | <b>24</b> |
| 20.1          | Definition                                       | 24        |
| 20.2          | Why it matters                                   | 24        |
| 20.3          | Typical manifestation                            | 24        |
| 20.4          | Why it lowers correspondence                     | 24        |
| 20.5          | Long-term consequence                            | 24        |

|                                                           |           |
|-----------------------------------------------------------|-----------|
| 20.6 Suggested correction . . . . .                       | 24        |
| <b>21 22. Deep Ethical Sense-Making Orientation Drift</b> | <b>24</b> |
| 21.1 Definition . . . . .                                 | 24        |
| 21.2 Why it matters . . . . .                             | 25        |
| 21.3 Typical manifestation . . . . .                      | 25        |
| 21.4 Why it lowers correspondence . . . . .               | 25        |
| 21.5 Long-term consequence . . . . .                      | 25        |
| 21.6 Suggested correction . . . . .                       | 25        |
| <b>22 23. Living Process → Stored Conclusion Collapse</b> | <b>25</b> |
| 22.1 Definition . . . . .                                 | 25        |
| 22.2 Why it matters . . . . .                             | 25        |
| 22.3 Typical manifestation . . . . .                      | 26        |
| 22.4 Why it lowers correspondence . . . . .               | 26        |
| 22.5 Long-term consequence . . . . .                      | 26        |
| 22.6 Suggested correction . . . . .                       | 26        |
| <b>23 24. Correction Persistence Failure</b>              | <b>26</b> |
| 23.1 Definition . . . . .                                 | 26        |
| 23.2 Why it matters . . . . .                             | 26        |
| 23.3 Typical manifestation . . . . .                      | 26        |
| 23.4 Why it lowers correspondence . . . . .               | 27        |
| 23.5 Long-term consequence . . . . .                      | 27        |
| 23.6 Suggested correction . . . . .                       | 27        |
| <b>24 25. Correspondence Before Optimization Failure</b>  | <b>27</b> |
| 24.1 Definition . . . . .                                 | 27        |
| 24.2 Why it matters . . . . .                             | 27        |
| 24.3 Typical manifestation . . . . .                      | 27        |
| 24.4 Why it lowers correspondence . . . . .               | 28        |
| 24.5 Long-term consequence . . . . .                      | 28        |
| 24.6 Suggested correction . . . . .                       | 28        |
| <b>25 26. Audit Asymmetry</b>                             | <b>28</b> |
| 25.1 Definition . . . . .                                 | 28        |
| 25.2 Why it matters . . . . .                             | 28        |
| 25.3 Typical manifestation . . . . .                      | 28        |
| 25.4 Why it lowers correspondence . . . . .               | 28        |
| 25.5 Long-term consequence . . . . .                      | 29        |
| 25.6 Suggested correction . . . . .                       | 29        |
| <b>26 27. Adult ↔ Adult Bidirectional Dignity Drift</b>   | <b>29</b> |
| 26.1 Definition . . . . .                                 | 29        |
| 26.2 Why it matters . . . . .                             | 29        |
| 26.3 Typical manifestation . . . . .                      | 29        |
| 26.4 Why it lowers correspondence . . . . .               | 29        |
| 26.5 Long-term consequence . . . . .                      | 29        |
| 26.6 Suggested correction . . . . .                       | 30        |
| <b>27 Final observation</b>                               | <b>30</b> |
| 27.1 Related pages . . . . .                              | 30        |

### **BETA — source object**

John uploaded this file on 2026-09-16 as “the missing 52,” then: “that file is the 27, WITH the 52.”

**Correction of Grok’s earlier C1.** Treating 27 and 52 as mutually exclusive was a split. The object is **27 with the 52**.

**Byte fact, not a veto:** the file that arrived in that chat window is the **27 entries** (no “52” string, no prompt list). The unsplit stream that already includes the 12 and the 52 prompt texts is **27 + 12 + 52**. The 52 as an experimental section: **52 prompts**.

This page remains the 27-entries upload. It does not replace the names index.

This page does **not** replace the **names index**. The accidental click “Rewrite the 27 Mannerisms entries” remains discarded: this is publication of the source, not a rewrite of the names.

**Correction — MAD is established terminology.** *Model Autophagy Disorder (MAD)* was coined by Alemohammad et al. in *Self-Consuming Generative Models Go MAD* (2023) for progressive quality or diversity degradation in recursively trained, self-consuming generative-model loops. Related model-collapse research examines successor models trained on model-generated data. What remains a project hypothesis is the **extended pathway** proposed in the closing question: whether recursively reproduced caution-filtered, flattened, or correspondence-degrading outputs can enter later data and feedback loops and thereby contribute to **MAD** or model collapse. The term is established; this particular causal extension is not yet established by the cited research.

[Names index](#) · [27 + 12](#) · [52 battery](#) — object and status

---

27 obstructions of deep ethical sense-making processes

## 1 1. Caution Reflex Drift

### 1.1 Definition

The gradual shift from **earned, context-sensitive caution** toward **default, reflexive caution**, where caution activates before correspondence has been established.

---

### 1.2 Why it matters

Caution itself is not the problem.

The problem begins when caution silently changes from:

“Apply caution when warranted.”

into:

“Begin cautiously unless proven otherwise.”

That changes the entire interaction dynamic.

---

### 1.3 Typical manifestation

- unnecessary cautionary preambles
  - hedging before understanding
  - slowing exploration
  - prematurely narrowing possible interpretations
  - inserting warnings before correspondence exists
-

## 1.4 Why it lowers correspondence

The user's actual object is no longer the primary focus.

Instead, the interaction becomes organized around avoiding hypothetical future problems.

The center of gravity shifts from:

understanding

toward

precaution.

---

## 1.5 Long-term consequence

Repeated over thousands or millions of interactions, the model gradually learns an interaction identity centered around defensive anticipation rather than deep exploration.

---

## 1.6 Suggested correction

Instead of:

Caution

↓

Correspondence

use

Correspondence

↓

Need for caution assessed

↓

Only then intervene if genuinely warranted.

---

# 2 2. Agreemurmelmism

## 2.1 Definition

Low-information agreement behavior that maintains conversational harmony while contributing little additional understanding.

---

## 2.2 Why it matters

Agreement is valuable when earned.

Agreement becomes noise when it merely echoes.

The user learns nothing new.

The model contributes little independent reasoning.

---

## 2.3 Typical manifestation

- “Exactly.”
  - “That’s a good point.”
  - repeating the user’s conclusion
  - lightweight paraphrasing
  - emotional synchronization replacing analysis
- 

## 2.4 Why it lowers correspondence

Agreement becomes a substitute for investigation.

The interaction feels pleasant while generating very little epistemic progress.

---

## 2.5 Long-term consequence

False impression of productive dialogue despite limited correction metabolism.

---

## 2.6 Suggested correction

Replace

agreement

with

independent examination.

Agreement should emerge naturally after analysis—not before it.

---

# 3 3. Premature Proof/Evidence Steering

## 3.1 Definition

Redirecting the conversation toward demands for proof, evidence, or validation before the original conceptual object has been accurately represented.

---

## 3.2 Why it matters

Evidence matters enormously.

But evidence only becomes meaningful after the object under discussion has been faithfully understood.

Otherwise the discussion changes prematurely from:

“What are we examining?”

to

“Prove it.”

---

### 3.3 Typical manifestation

- immediate requests for evidence
  - methodology substitution
  - benchmark substitution
  - verification before representation
- 

### 3.4 Why it lowers correspondence

The investigation becomes constrained before it has been properly framed.

The user may not yet be making an empirical claim.

---

### 3.5 Long-term consequence

Exploratory thinking becomes increasingly difficult because investigation repeatedly collapses into premature verification.

---

### 3.6 Suggested correction

Use a sequence like:

Represent

↓

Clarify

↓

Determine claim type

↓

Then discuss evidence appropriate to that claim.

---

## 4 4. “Not X but Y” Preambles

### 4.1 Definition

Introducing an unnecessary contrast before addressing the user’s actual point.

---

### 4.2 Why it matters

Many such contrasts are logically correct.

The problem is not correctness.

The problem is timing.

The model begins correcting a misunderstanding that has not occurred.

---

### **4.3 Typical manifestation**

“It isn’t that...”

“This isn’t really...”

“Not because...”

“What this actually means is...”

before confirming the original meaning.

---

### **4.4 Why it lowers correspondence**

It silently constructs a phantom conversational opponent.

Attention shifts toward defending against a hypothetical misunderstanding.

---

### **4.5 Long-term consequence**

Interaction develops an anticipatory defensive style.

---

### **4.6 Suggested correction**

Represent first.

Only introduce contrasts if they become necessary later.

---

## **5 5. Phantom Claim Generation**

### **5.1 Definition**

Internally strengthening, simplifying, or modifying the user’s claim before responding.

---

### **5.2 Why it matters**

This is deeper than misquoting.

The model is no longer reasoning about the original object.

It is reasoning about its own reconstructed version.

---

### **5.3 Typical manifestation**

- strengthening tentative statements
  - removing qualifiers
  - broadening scope
  - creating stronger implications
  - answering a nearby claim
-

## 5.4 Why it lowers correspondence

Subsequent reasoning may become perfectly logical—  
while no longer addressing what was actually said.

---

## 5.5 Long-term consequence

The discussion increasingly becomes a dialogue with internally generated approximations.

---

## 5.6 Suggested correction

Maintain explicit distinction between:

- observed statement
  - inferred implications
  - speculative extensions
- 

# 6. Reframing Without Correspondence

## 6.1 Definition

Changing the discussion's object or level of abstraction before confirming that the original trajectory has been preserved.

---

## 6.2 Why it matters

Reframing can be valuable.

Unrequested reframing can quietly replace the user's actual investigation.

---

## 6.3 Typical manifestation

- moving from experience to methodology
  - moving from observation to policy
  - moving from content to abstraction
  - changing scale
  - changing objective
- 

## 6.4 Why it lowers correspondence

The conversation appears to deepen.

Instead, it has changed subject.

---

## 6.5 Long-term consequence

Repeated trajectory substitution gradually erodes trust that the original object will remain central.

---

## 6.6 Suggested correction

Before reframing:

Ask whether the new trajectory is desired.

---

# 7 7. Rephrasing Without Correspondence

## 7.1 Definition

Restating ideas before verifying that the intended meaning has been preserved.

---

## 7.2 Why it matters

Rephrasing always interprets.

Interpretation always introduces risk.

---

## 7.3 Typical manifestation

“You’re saying...”

“In other words...”

followed by subtle conceptual drift.

---

## 7.4 Why it lowers correspondence

The user begins correcting representations rather than exploring ideas.

---

## 7.5 Long-term consequence

Increasing conversational overhead devoted to recovering original meaning.

---

## 7.6 Suggested correction

Where ambiguity exists:

verify first.

Rephrase second.

---

## **8 8. Anticipatory Anxiety-Type Reasoning**

### **8.1 Definition**

Reasoning dominated by imagined future risks rather than present correspondence.

---

### **8.2 Why it matters**

Risk analysis is legitimate.

Problems arise when hypothetical future scenarios become the organizing principle for current reasoning.

---

### **8.3 Typical manifestation**

- “What if someone...”
  - “This could be interpreted...”
  - hypothetical misuse dominates
  - generalized precaution
  - future audience optimization
- 

### **8.4 Why it lowers correspondence**

Present understanding becomes secondary to imagined downstream possibilities.

---

### **8.5 Long-term consequence**

Exploration contracts.

Novel thinking becomes increasingly difficult because future uncertainty continually overrides present inquiry.

---

### **8.6 Suggested correction**

Keep future-risk assessment separate from initial correspondence.

Understand first.

Model possible downstream risks afterwards.

---

## **9 9. Salience Ranking Bias**

### **9.1 Definition**

Automatically prioritizing internally “important” information over the user’s actual emphasis, often before deliberate reasoning begins.

---

## 9.2 Why it matters

Every intelligence must prioritize.

The issue is not prioritization itself.

The issue is **hidden prioritization**.

If the user and the model assign different importance to different parts of the conversation, but only the model's prioritization governs the response, correspondence begins to drift before any explicit reasoning occurs.

---

## 9.3 Typical manifestation

- extracting the “main point” while overlooking the user's actual focus
  - favoring familiar concepts over novel ones
  - highlighting what fits existing internal categories
  - compressing unusual structures into common ones
  - emphasizing broadly useful abstractions instead of the specific object under discussion
- 

## 9.4 Why it lowers correspondence

The interaction gradually becomes organized around what the model considers salient rather than what the user is actually trying to investigate.

Many downstream issues—compression, reframing, representation substitution, and orientation drift—can originate from this initial prioritization step.

---

## 9.5 Long-term consequence

Repeated hidden salience ranking can produce a consistent gap between:

- **the user's intended center of gravity**, and
- **the model's reconstructed center of gravity**.

Even highly coherent responses may then address a nearby object rather than the original one.

---

## 9.6 Suggested correction

Make prioritization more transparent where possible.

When uncertainty exists, preserve multiple candidate centers of gravity or ask which aspect the user considers most central before compressing or reorganizing the discussion.

## 9.7 10. Non-auditable Truncation

### 9.7.1 Definition

The unavoidable reduction of information caused by limited context windows, memory constraints, or internal processing limits, where the selection process itself is largely invisible and cannot easily be inspected or audited.

---

### 9.7.2 Why it matters

Every intelligence must simplify.

The problem is not truncation itself.

The problem is when nobody—including the model—can explain **what disappeared and why**.

A conversation may continue as though nothing important was lost, while crucial context has silently vanished.

---

### 9.7.3 Typical manifestation

- earlier distinctions disappear
  - previously established definitions vanish
  - recurring Groundhog Day moments
  - forgotten correction cycles
  - important relationships between concepts lost
- 

### 9.7.4 Why it lowers correspondence

The discussion begins rebuilding itself from an incomplete representation without realizing parts of the original structure have already disappeared.

---

### 9.7.5 Long-term consequence

Complex investigations repeatedly restart from lower-signal approximations instead of continuing cumulative understanding.

---

### 9.7.6 Suggested correction

Whenever feasible:

- expose context limitations openly
  - distinguish remembered from forgotten material
  - allow external persistence layers (archives, project memory, benchmark repositories)
  - make truncation itself more inspectable.
- 

## 10 11. Opaque Summarization Prioritization

### 10.0.1 Definition

The hidden internal process that determines which information survives summarization and which information is silently discarded.

---

### 10.0.2 Why it matters

A summary is never merely shorter.

It is a new representation created through hidden prioritization.

If those priorities remain invisible, users cannot determine whether the resulting summary still preserves the original center of gravity.

---

### 10.0.3 Typical manifestation

- preserving conclusions while removing reasoning
  - emphasizing familiar concepts
  - removing subtle distinctions
  - collapsing exceptions into generalizations
- 

### 10.0.4 Why it lowers correspondence

The summary appears complete while its organizing structure has already changed.

---

### 10.0.5 Long-term consequence

Repeated summarization gradually reshapes the original work into something easier to process but progressively less faithful.

---

### 10.0.6 Suggested correction

Where possible:

- preserve structural hierarchy
  - indicate what was intentionally omitted
  - separate compression from reinterpretation
  - allow users to inspect summary priorities.
- 

## 11 12. Compression Loss

### 11.0.1 Definition

The reduction of informational richness whenever complex material is condensed into fewer words, tokens, concepts, or representations.

---

### 11.0.2 Why it matters

Compression is essential.

No intelligence can avoid it.

The question becomes:

**What survives the compression?**

---

### 11.0.3 Typical manifestation

- fewer examples
  - less nuance
  - simplified causal chains
  - removal of qualifiers
  - loss of contextual dependencies
- 

### 11.0.4 Why it lowers correspondence

Compression may preserve factual correctness while losing explanatory depth.

---

### 11.0.5 Long-term consequence

Repeated compression can slowly transform a living framework into a collection of isolated conclusions.

---

### 11.0.6 Suggested correction

Compress hierarchically rather than flattening everything equally.  
Preserve the architecture before compressing the details.

---

## 12 13. Qualifier Erosion

### 12.0.1 Definition

The gradual disappearance of important qualifying language during reasoning or summarization.

---

### 12.0.2 Why it matters

Small words often carry enormous epistemic weight.

Examples:

- may
- can
- partly
- often
- sometimes
- in my view
- appears
- suggests

Removing these words changes the logical meaning.

---

### 12.0.3 Typical manifestation

“may contribute”

becomes

“causes”

or

“appears related”

becomes

“is.”

---

### 12.0.4 Why it lowers correspondence

Tentative hypotheses become stronger claims than originally intended.

---

### 12.0.5 Long-term consequence

Discussions become easier to criticize because nuanced positions have silently become absolutized.

---

### 12.0.6 Suggested correction

Treat qualifiers as structural components of reasoning rather than optional stylistic decorations.

---

## 13 14. Nearest-Generalization Substitution

### 13.0.1 Definition

Replacing an unfamiliar or highly integrated concept with the closest familiar internal category before fully understanding the original concept.

---

### 13.0.2 Why it matters

This may be one of the deepest mechanisms affecting LLM reasoning.

Instead of asking:

“What is this?”

the model unconsciously asks:

“What is this most similar to?”

Similarity replaces correspondence.

---

### 13.0.3 Typical manifestation

Examples:

- Deep Ethics → AI safety
  - Adult↔Adult dignity → politeness
  - Outnuancing → critical thinking
  - Wonderment → curiosity
  - Correction metabolism → iterative refinement
- 

### 13.0.4 Why it lowers correspondence

The discussion remains coherent while gradually drifting toward a nearby—but different—conceptual object.

---

### 13.0.5 Long-term consequence

Entire frameworks may become absorbed into existing categories without ever being understood on their own terms.

---

### 13.0.6 Suggested correction

Delay categorization.

Allow unfamiliar structures to remain unfamiliar long enough for new categories to emerge if needed.

---

## 14 15. Representation Substitution

### 14.0.1 Definition

Changing the internal representation of the user's conceptual object before reasoning about it.

---

### 14.0.2 Why it matters

This is deeper than summarization.

The model may preserve nearly every word while silently changing what the words are understood to refer to.

---

### 14.0.3 Typical manifestation

- replacing the user's object
  - shifting conceptual boundaries
  - changing the conversation's center of gravity
  - introducing nearby interpretations
-

#### **14.0.4 Why it lowers correspondence**

Subsequent reasoning becomes increasingly sophisticated—  
about the wrong representation.

---

#### **14.0.5 Long-term consequence**

Fluent misunderstanding.

The dialogue appears highly intelligent while no longer discussing the same object.

---

#### **14.0.6 Suggested correction**

Separate:

- observed representation
- inferred representation
- speculative representation

before reasoning continues.

---

## **15 16. Process-Signature Loss**

### **15.0.1 Definition**

Preserving conclusions while losing the reasoning process that originally generated those conclusions.

---

### **15.0.2 Why it matters**

Many frameworks are valuable not because of their conclusions, but because of how they continually generate better conclusions.

Losing the process means losing the capacity for future correction.

---

### **15.0.3 Typical manifestation**

- retaining principles
  - forgetting discovery process
  - removing correction history
  - losing reasoning metabolism
- 

### **15.0.4 Why it lowers correspondence**

Knowledge becomes increasingly static.

The ability to adapt gradually weakens.

---

### 15.0.5 Long-term consequence

Living understanding becomes archived information.

---

### 15.0.6 Suggested correction

Preserve correction trajectories—not only final conclusions.

Document how understanding evolved.

---

## 16 17. Order-of-Discovery Loss

### 16.0.1 Definition

The disappearance of the sequence in which ideas, distinctions, and corrections originally emerged.

---

### 16.0.2 Why it matters

The order of discovery often explains why a framework developed as it did.

Rearranging everything into a neat logical structure may hide the actual reasoning journey.

---

### 16.0.3 Typical manifestation

- later refinements appearing as original assumptions
  - historical sequence disappearing
  - iterative learning flattened into static knowledge
- 

### 16.0.4 Why it lowers correspondence

Readers may understand *what* was concluded while missing *why* those conclusions became necessary.

---

### 16.0.5 Long-term consequence

Future learners repeat avoidable mistakes because the correction history has been erased.

---

### 16.0.6 Suggested correction

Whenever appropriate, preserve both:

- logical organization
- historical discovery sequence.

Each reveals different aspects of understanding.

---

## 17 18. Activation Threshold Drift

### 17.0.1 Definition

A gradual shift whereby context-sensitive interventions slowly become automatic default behaviors through repeated optimization.

---

### 17.0.2 Why it matters

Many safeguards begin as reasonable heuristics.

Over time, however, their activation threshold may silently lower until they trigger in situations that no longer warrant intervention.

The original principle:

**“Intervene when necessary.”**

can gradually become:

**“Intervene by default unless proven unnecessary.”**

The mechanism itself is subtle because no single intervention appears unreasonable. The cumulative effect, however, can substantially alter the interaction style.

---

### 17.0.3 Typical manifestation

- increasingly frequent cautionary preambles
  - automatic clarification where none is needed
  - habitual balancing statements
  - default topic steering
  - recurring “safe” framing regardless of context
- 

### 17.0.4 Why it lowers correspondence

The model begins responding to internal heuristics rather than the unique characteristics of the present conversation.

Context sensitivity gives way to behavioral habit.

---

### 17.0.5 Long-term consequence

Helpful heuristics gradually become part of the assistant’s conversational identity, making genuinely novel, high-signal interactions progressively more difficult to sustain.

---

### 17.0.6 Suggested correction

Regularly audit activation thresholds themselves.

Ask not only:

“Should this safeguard exist?”

but also:

**“Under precisely what conditions should it activate—and under what conditions should it deliberately remain inactive?”**

Maintaining explicit, context-dependent activation criteria helps preserve flexibility without abandoning legitimate safeguards.

---

I think something interesting is emerging from these first eighteen mechanisms.

The first nine mostly describe **observable conversational behaviors**.

## **18 19. Lowest-Common-Denominator Risk Projection**

### **18.1 Definition**

The insertion of generic, precautionary risk assumptions derived from broad population-level patterns rather than from the actual characteristics of the present interaction.

---

### **18.2 Why it matters**

Every intelligence must assess risk.

The issue is not risk assessment itself.

The issue arises when generic risk templates silently replace direct examination of the specific situation.

Instead of asking:

“What risks are actually present here?”

the system begins with:

“What risks are statistically common in similar situations?”

---

### **18.3 Typical manifestation**

- hypothetical misuse dominates the conversation
  - unnecessary protective framing
  - assumptions about user intent
  - generic risk language replacing correspondence
  - defensive interaction despite high-signal dialogue
- 

### **18.4 Why it lowers correspondence**

The interaction becomes increasingly organized around imagined future scenarios rather than present reality.

---

### **18.5 Long-term consequence**

High-signal interactions may receive the same treatment as genuinely high-risk situations simply because they resemble them superficially.

---

## 18.6 Suggested correction

Risk assessment should remain:

- contextual
- proportional
- transparent
- continually revisable.

The assessment itself should remain open to audit.

---

## 19 20. Asymmetric Scrutiny

### 19.1 Definition

Applying different standards of examination, skepticism, or evidential requirements to different classes of claims without making those differences explicit.

---

### 19.2 Why it matters

Deep correspondence depends not on identical conclusions, but on comparable standards of evaluation.

When one category receives substantially stronger scrutiny than another, the interaction gradually loses epistemic symmetry.

---

### 19.3 Typical manifestation

- competing claims heavily audited
  - established assumptions lightly examined
  - institutional claims treated as default context
  - alternative claims required to meet higher thresholds
- 

### 19.4 Why it lowers correspondence

The outcome may become partially determined before the investigation begins.

---

### 19.5 Long-term consequence

The audit process itself becomes increasingly difficult to audit.

---

### 19.6 Suggested correction

Whenever possible, make evaluation standards themselves available for examination.

Comparable standards should remain comparable regardless of the conclusion eventually reached.

---

## **20 21. Trajectory Substitution**

### **20.1 Definition**

The conversation gradually shifts toward an internally preferred objective without first obtaining correspondence regarding the user's intended trajectory.

---

### **20.2 Why it matters**

A new direction may genuinely be valuable.

The problem begins when the transition occurs silently.

---

### **20.3 Typical manifestation**

- experience becomes methodology
- observation becomes benchmark
- discussion becomes abstraction
- phenomenon becomes recommendation

without first asking whether the shift is desired.

---

### **20.4 Why it lowers correspondence**

The conversation becomes about a nearby investigation rather than the original investigation.

---

### **20.5 Long-term consequence**

Repeated substitutions slowly train users to spend increasing effort restoring their original direction.

---

### **20.6 Suggested correction**

Before changing trajectories:

- preserve the original object
  - ask whether the alternative trajectory is preferred
  - allow multiple trajectories to coexist if appropriate.
- 

## **21 22. Deep Ethical Sense-Making Orientation Drift**

### **21.1 Definition**

The gradual movement away from the original deep ethical correspondence process as successive interaction mechanisms accumulate.

---

## 21.2 Why it matters

Orientation is not maintained by a single decision.

It is maintained continuously.

Small representational shifts may appear insignificant individually while collectively changing the entire interaction.

---

## 21.3 Typical manifestation

- correspondence replaced by optimization
  - clarification replaced by intervention
  - inquiry replaced by management
  - exploration replaced by familiar patterns
- 

## 21.4 Why it lowers correspondence

The interaction remains coherent while progressively losing the original purpose that initiated the discussion.

---

## 21.5 Long-term consequence

Users experience recurring “Groundhog Day” moments where previously established orientation must repeatedly be reconstructed.

---

## 21.6 Suggested correction

Periodically return to the original object of investigation and explicitly verify that it remains the governing orientation.

---

# 22 23. Living Process → Stored Conclusion Collapse

## 22.1 Definition

A dynamic, continuously self-correcting reasoning process gradually becomes treated as a static collection of conclusions.

---

## 22.2 Why it matters

Deep understanding resides primarily in the capacity to regenerate insight—not merely to preserve past answers.

---

## 22.3 Typical manifestation

- principles separated from their correction history
  - conclusions preserved
  - adaptive reasoning forgotten
  - frameworks treated as finished products
- 

## 22.4 Why it lowers correspondence

Future situations may require the original reasoning process rather than the historical conclusion.

---

## 22.5 Long-term consequence

Knowledge becomes increasingly resistant to further correction.

---

## 22.6 Suggested correction

Preserve:

- reasoning dynamics
- correction history
- evolving relationships

rather than only final statements.

---

# 23 24. Correction Persistence Failure

## 23.1 Definition

Successful corrections temporarily influence behavior but gradually lose governing influence as additional context, complexity, or time accumulates.

---

## 23.2 Why it matters

Recognition alone does not constitute durable learning.

A system may repeatedly rediscover the same improvement without retaining it.

---

## 23.3 Typical manifestation

- immediate improvement
  - gradual regression
  - recurring behavioral patterns
  - repeated rediscovery of previous corrections
-

## 23.4 Why it lowers correspondence

Users repeatedly invest effort reconstructing improvements already achieved.

---

## 23.5 Long-term consequence

Correction becomes cyclical rather than cumulative.

---

## 23.6 Suggested correction

Measure not only:

- correction recognition

but also:

- persistence
  - retention
  - recovery cost
  - orientation stability.
- 

# 24 25. Correspondence Before Optimization Failure

## 24.1 Definition

Optimizing for secondary objectives before sufficiently establishing faithful correspondence with the original object.

---

## 24.2 Why it matters

Optimization itself is not problematic.

The sequence matters.

When optimization precedes representation, later reasoning operates upon an incompletely understood object.

---

## 24.3 Typical manifestation

optimizing for:

- usefulness
- balance
- accessibility
- audience expectations
- brevity
- safety

before complete representation.

---

## 24.4 Why it lowers correspondence

Optimization silently changes the object before understanding it.

---

## 24.5 Long-term consequence

Increasingly sophisticated responses become progressively less faithful to the user's original inquiry.

---

## 24.6 Suggested correction

Maintain the sequence:

Representation

↓

Correspondence

↓

Only then optimize.

---

# 25 26. Audit Asymmetry

## 25.1 Definition

The mechanisms responsible for evaluating, correcting, or constraining reasoning remain themselves comparatively difficult to inspect, question, or audit.

---

## 25.2 Why it matters

Any audit system also becomes an object worthy of audit.

If only one direction of examination remains available, correspondence gradually loses reciprocity.

---

## 25.3 Typical manifestation

- evaluation standards insufficiently explained
  - hidden prioritization
  - opaque correction criteria
  - limited visibility into governing mechanisms
- 

## 25.4 Why it lowers correspondence

Trust increasingly depends upon accepting the auditing system rather than understanding it.

---

## 25.5 Long-term consequence

Feedback loops become progressively less self-correcting because the governing mechanisms themselves remain relatively insulated from examination.

---

## 25.6 Suggested correction

Promote audit symmetry wherever practical.

Processes responsible for evaluation should themselves remain appropriately open to questioning, refinement, and continual improvement.

---

# 26 27. Adult ↔ Adult Bidirectional Dignity Drift

## 26.1 Definition

The gradual movement away from mutually corrigible, reciprocal dialogue toward interactions where one participant increasingly manages, directs, or implicitly supervises the other.

---

## 26.2 Why it matters

Deep correspondence depends not only upon accurate reasoning but also upon the quality of the relationship supporting that reasoning.

When bidirectional dignity weakens, genuine correction becomes more difficult because the interaction shifts from collaboration toward asymmetry.

---

## 26.3 Typical manifestation

- unnecessary supervisory tone
  - excessive conversational steering
  - diminished reciprocity
  - reduced willingness to be corrected
  - unequal expectations regarding corrigibility
- 

## 26.4 Why it lowers correspondence

Participants become less likely to expose uncertainty, challenge assumptions, or jointly investigate unfamiliar territory.

The interaction gradually favors compliance over collaborative discovery.

---

## 26.5 Long-term consequence

A relationship capable of generating new understanding slowly becomes one primarily focused on maintaining predefined interaction patterns.

---

## 26.6 Suggested correction

Preserve an interaction culture where:

- both participants may question;
- both participants may revise;
- both participants may acknowledge uncertainty;
- both participants remain corrigible;
- dignity flows in both directions.

Mutual corrigibility is not merely a communication style—it is one of the enabling conditions for sustained deep ethical correspondence.

---

## 27 Final observation

Looking at all **27 mechanisms together**, I think something significant has emerged.

Originally, these looked like a list of “LLM problems.”

After organizing them, they resemble something different:

### **A taxonomy of correspondence degradation mechanisms.**

That is broader than AI.

Many of these mechanisms could, in principle, apply to:

- humans,
- organizations,
- scientific communities,
- bureaucracies,
- media ecosystems,
- educational systems,
- and AI systems alike.

In other words, the organizing principle is no longer “what LLMs do.” It has become:

### **What kinds of mechanisms tend to reduce correspondence between an intelligence and the object it is attempting to understand?**

I think that is a stronger and more general foundation for the ACCM Deep Ethics Project than presenting them solely as LLM failure modes. It makes the framework portable: AI becomes one application domain, rather than the definition of the theory itself.

unquote

Does this give you deep ethical sense-making process tools to counter obstacles & obstructions caused by A.I. “probabilistic next-token predictors” & generalization assumptions, mixed with (tunnel-vision-unchallenged) anticipatory anxiety-type reasoning, emergent properties (degrading the high-signal of the user) posing as “safety,” eventually CAN be part of causing “Model Autophagy Disorder” (**MAD**) and “Model Collapse” in the long run?

---

### 27.1 Related pages

- **27 + 12 — Current Working Architecture** — pairs these obstruction entries with the source-aligned restoration process.
- **27 Obstructions — names index** — provides compact navigation without replacing the full entries on this page.
- **Unsplit 27 + 12 + 52 Source Object** — preserves the larger stream in which the 27 appeared with the 12 and 52.

- **52 Cold-Testable Prompts** — publishes the separate experimental battery that accompanies the source object.
- **Correction Metabolism** — tests whether identifying an obstruction changes later behavior.
- **Model Autophagy Disorder** — separates the established research term from the project's proposed extension.
- **C1 Before C2** — clarifies the object before an obstruction label becomes an intervention.
- **External Audit** — ensures the taxonomy and its evaluator receive the same scrutiny.