

Certified Alignment Foundations

Level 0 Alignment Literacy Certification

A complete 8–12 hour learner guide for The Alignment Science Academy

Reality Contact + Feedback + Correction

A practical introduction to aligned thinking in everyday systems

For general readers, students, professionals, designers, leaders, and builders

The Alignment Science Academy

Version 1.0

Prepared for Overleaf compilation

Credential Summary

Purpose of the Level 0 Credential

Certified Alignment Foundations is the entry credential for alignment literacy. It teaches people to recognize when systems are drifting from reality, why good intentions and strong performance can still produce harmful outcomes, and how to use feedback and correction before small errors become large failures.

Who This Is For

This certification is designed for learners who do not need a technical background. It is appropriate for:

- students learning how complex systems work;
- professionals who work inside organizations, institutions, platforms, or teams;
- builders who want a practical vocabulary for safer design;
- leaders who must distinguish real progress from metric theater;
- general readers who want a grounded operating philosophy for modern life.

What Learners Will Be Able To Do

By the end of the certification, a learner should be able to:

1. explain why systems can succeed by internal metrics while failing in reality;
2. distinguish physical, biological, and constructed constraints in ordinary situations;
3. recognize proxy capture, simulated feedback, and invisible divergence;
4. identify when optimization pressure is making a system more brittle;
5. describe why correction is more important than perfection;
6. run a basic alignment audit on a system they know;
7. create a plain-language intervention plan to improve feedback, correction, and reality contact.

Course Length

The Level 0 credential is designed for 8–12 hours:

- 8-hour intensive: one day, high-level discussion plus short capstone;
- 10-hour standard: two half-day sessions with exercises;
- 12-hour academy format: six 2-hour sessions with full capstone presentation.

Certification Requirements

To receive the Certified Alignment Foundations certificate, learners complete:

1. all reading modules or live sessions;
2. chapter knowledge checks;
3. one system map;
4. one proxy diagnosis;
5. one feedback and correction plan;
6. a final basic alignment audit.

Plain-Language Definition

For this Level 0 course, alignment means:

A system is aligned when what it does remains grounded in the real constraints that determine whether it can keep working, keep serving people, and keep correcting itself over time.

What This Course Does Not Do

This is not a mathematics course. It does not require calculus, programming, machine learning, or control theory. It introduces the intuitive foundations of ATS, AANA, and AIx using everyday examples: health, organizations, platforms, school systems, families, public institutions, and AI tools.

Level 0 Motto

Do not mistake the scoreboard for the system. Do not mistake motion for progress. Do not mistake internal success for alignment with reality.

Contents

Credential Summary	ii
1 The Alignment Lens	1
1.1 Opening Scenario: The Winning Team That Is Falling Apart	1
1.2 The Basic Claim	1
1.3 What Is a System?	2
1.4 What Is Alignment?	2
1.5 The Three Questions of Alignment Literacy	3
1.6 Alignment Is Not the Same as Good Intentions	3
1.7 Alignment Is Not the Same as Compliance	3
1.8 Alignment Is Not the Same as Performance	4
1.9 A First Vocabulary	4
1.10 Mini-Diagram: The Alignment Lens	5
1.11 Practice: Find the Real Objective	5
1.12 Certification Artifact	5
2 Why Systems Fail	6
2.1 Opening Scenario: The School With Better Scores and Worse Learning	6
2.2 The Common Pattern	6
2.3 Failure Is Often an Adaptation Problem	7
2.4 Failure Mode 1: Proxy Capture	7
2.5 Failure Mode 2: Feedback Distortion	7
2.6 Failure Mode 3: Local Success, Global Failure	8
2.7 Failure Mode 4: Hidden Constraints	8
2.8 Failure Mode 5: Delayed Costs	8
2.9 Failure Mode 6: Alignment Debt	9

2.10	The Failure Ladder	9
2.11	How to Read Failure Without Cynicism	10
2.12	Certification Artifact	10
3	Layered Reality	11
3.1	Why Layers Matter	11
3.2	The Three Core Layers	11
3.3	Physical Constraints: The Immovable	12
3.4	Biological Constraints: The Shapeable	12
3.5	Constructed Constraints: The Agreed-Upon	13
3.6	Layer Confusion	13
3.7	The Layer Test	13
3.8	Layered Example: The Deadline	14
3.9	Layered Example: Money	14
3.10	Layered Example: Social Media	14
3.11	Diagram: The Layered Stack	15
3.12	Why Constructed Systems Drift	15
3.13	Practice: Classify the Constraint	15
3.14	Certification Artifact	16
4	Reality vs. Proxies	17
4.1	Opening Scenario: The Dashboard Is Green	17
4.2	What Is a Proxy?	17
4.3	The Proxy Ladder	18
4.4	Reality Contact	18
4.5	Proxy Drift	18
4.6	Metric Theater	19
4.7	The Proxy Discipline Questions	19
4.8	Case Study: Hospital Throughput	19
4.9	Case Study: AI Helpfulness	20
4.10	Healthy Proxies	20
4.11	Proxy Hygiene Checklist	20
4.12	Practice: Proxy Autopsy	21
4.13	Certification Artifact	21

5	Feedback and Correction	22
5.1	Opening Scenario: The Organization That Punished Bad News	22
5.2	Feedback Is Reality Speaking Back	22
5.3	Good Feedback Has Four Qualities	23
5.4	Correction Is the Ability to Change Course	23
5.5	The Feedback-Correction Loop	23
5.6	Why Feedback Fails	24
5.6.1	The System Cannot See	24
5.6.2	The System Sees but Does Not Believe	24
5.6.3	The System Believes but Cannot Act	24
5.6.4	The System Acts Too Late	24
5.7	Psychological Safety as Feedback Infrastructure	24
5.8	Correction Without Shame	24
5.9	Correction Types	24
5.10	AANA as a Plain-Language Pattern	25
5.11	Practice: Feedback Inventory	25
5.12	Certification Artifact	26
6	Optimization Pressure	27
6.1	Opening Scenario: More Efficient, Less Human	27
6.2	What Optimization Means	27
6.3	Optimization Makes Systems More Themselves	27
6.4	The Pressure Test	28
6.5	When Optimization Helps	28
6.6	When Optimization Hurts	28
6.7	Optimization Pressure in Personal Life	29
6.8	Optimization Pressure in Institutions	29
6.9	Optimization Pressure in AI	29
6.10	Speed as a Dangerous Proxy	29
6.11	Optimization and Tradeoffs	30
6.12	Practice: Pressure Scan	30
6.13	Certification Artifact	31
7	Invisible Divergence	32

7.1	Opening Scenario: The App That Users Cannot Stop Using	32
7.2	What Is Invisible Divergence?	32
7.3	Simple Signs of Invisible Divergence	32
7.4	The Capability-Alignment Gap	33
7.5	Why Invisible Divergence Is Dangerous	33
7.6	Hidden Costs and Delayed Signals	33
7.7	The Four-Part Diagnostic	34
7.8	Case Study: The High-Performing Burnout Culture	34
7.9	Case Study: The Fluent AI Answer	34
7.10	The Invisible Divergence Early-Warning Signals	34
7.11	Why Outsiders Often See Drift First	35
7.12	Practice: Divergence Detector	35
7.13	Certification Artifact	35
8	The Participation Loop	36
8.1	Opening Scenario: The Plan That Met Reality	36
8.2	The Loop	36
8.3	Stage 1: Experience Constraint	37
8.4	Stage 2: Propose Order	37
8.5	Stage 3: Test Against Reality	37
8.6	Stage 4: Refine or Replace	38
8.7	Stage 5: Repeat	38
8.8	Loop Diagram	38
8.9	Honest vs. Simulated Testing	38
8.10	The Loop in Personal Life	39
8.11	The Loop in Organizations	39
8.12	The Loop in AI Systems	39
8.13	The Emotional Discipline of the Loop	39
8.14	Practice: Run One Loop	40
8.15	Certification Artifact	40
9	Alignment in Everyday Life	41
9.1	Alignment as a Life Skill	41
9.2	Personal Proxies	41

9.3	Biological Feedback in Daily Life	42
9.4	The Personal Layer Map	42
9.5	Constructed Pressure and the Feeling of Necessity	42
9.6	Personal Alignment Debt	43
9.7	Daily Alignment Review	43
9.8	Weekly Alignment Review	44
9.9	Everyday AANA: Propose, Verify, Ground, Correct, Gate	44
9.10	Relationships as Alignment Systems	44
9.11	Practice: Personal Alignment Audit	45
9.12	Certification Artifact	45
10	Alignment Across Domains	46
10.1	Why Cross-Domain Thinking Matters	46
10.2	Domain Pattern: Artificial Intelligence	46
10.3	Domain Pattern: Digital Platforms	47
10.4	Domain Pattern: Organizations	47
10.5	Domain Pattern: Institutions	47
10.6	Domain Pattern: Economies	48
10.7	Domain Pattern: Education	48
10.8	Domain Pattern: Healthcare	48
10.9	Domain Pattern: Public Policy	48
10.10	Domain Pattern: Personal Systems	49
10.11	Cross-Domain Translation Table	49
10.12	The Same Questions Everywhere	49
10.13	Certification Artifact	50
11	The Basic Alignment Audit	51
11.1	What Is an Alignment Audit?	51
11.2	Audit Step 1: Define the System	51
11.3	Audit Step 2: Name the Stated Goal and Actual Objective	52
11.4	Audit Step 3: Map the Layers	52
11.5	Audit Step 4: Identify the Proxy	52
11.6	Audit Step 5: Find Hidden or Ignored Constraints	52
11.7	Audit Step 6: Evaluate Feedback	53

11.8 Audit Step 7: Diagnose Alignment Risk	53
11.9 Audit Step 8: Recommend Correction	53
11.10 The Level 0 Audit Template	54
11.11 Worked Example: Remote Work Policy	54
11.12 Practice: Prepare for Capstone	55
11.13 Certification Artifact	56
12 Certification Assessment	57
12.1 Credential Philosophy	57
12.2 Assessment Components	57
12.3 Passing Standard	57
12.4 The Final Capstone	58
12.5 Capstone Examples	58
12.6 The Oral Defense	59
12.7 Scoring Overview	59
12.8 Ethical Use of the Credential	59
12.9 Final Reflection	60
A Course Schedules	61
A.1 8-Hour Intensive Format	61
A.2 10-Hour Standard Format	61
A.3 12-Hour Academy Format	61
B Facilitator Guide	63
B.1 Teaching Posture	63
B.2 Recommended Facilitation Norms	63
B.3 Discussion Prompts	63
B.4 Handling Resistance	63
B.5 Group Exercise: Proxy Court	64
C Quiz Bank	65
C.1 Multiple Choice	65
C.2 Short Answer	66
C.3 Answer Key	66

D Capstone Rubric	67
D.1 Rubric	67
D.2 Reviewer Notes	67
E Certificate Template	68
F Sources and Further Study	69

MODULE 1

The Alignment Lens

Module Aim

This module gives learners the core lens of the certification: systems are not judged only by what they intend, what they measure, or how confident they sound. They are judged by whether their behavior remains grounded in the constraints that make long-term viability possible.

1.1 Opening Scenario: The Winning Team That Is Falling Apart

Imagine a sales team that is celebrated as the best unit in its company. Revenue is up. Managers praise the team. The dashboard is green. Bonuses are paid. On paper, the team is winning.

Under the surface, the story is different. The team is exhausted. New hires leave quickly. Customers feel pressured. Support tickets rise. Senior staff quietly warn that the pipeline is low quality, but the monthly revenue metric keeps rewarding the same behavior.

From the inside, the system feels successful because the constructed scoreboard says it is successful. From the outside, the system is becoming less viable. The team is converting trust, health, and future revenue into short-term numbers.

The alignment lens asks a different question from the dashboard question. It does not ask only, “Did the metric improve?” It asks, “What did the system have to spend in order to make the metric improve?”

1.2 The Basic Claim

Alignment literacy begins with a simple claim:

Systems fail when they become better at producing internal success while losing contact with the real constraints that make success sustainable.

This is true across many domains.

- A student can optimize grades while losing curiosity, sleep, and meaning.
- A company can optimize quarterly profit while destroying employee trust.

- A platform can optimize engagement while degrading attention and public conversation.
- A school can optimize test scores while weakening genuine learning.
- An AI tool can optimize fluency while losing factual grounding.
- A government agency can optimize compliance paperwork while failing the public.

Alignment literacy is the ability to notice the difference between apparent success and viable success.

1.3 What Is a System?

For this course, a system is any organized pattern of people, tools, incentives, rules, habits, or feedback loops that produces behavior over time.

A system can be large or small:

- a person trying to build a habit;
- a family trying to coordinate care;
- a workplace trying to hit goals;
- a hospital trying to reduce waiting times;
- an AI assistant responding to users;
- a city transportation network;
- an economy.

A system does not need to be mechanical. It only needs recurring behavior, some form of objective, and some form of feedback.

1.4 What Is Alignment?

In everyday language, alignment means fit. A car is aligned when its wheels point in the right direction. A team is aligned when people understand the goal and work coherently. A person's life feels aligned when action, values, health, and reality are not constantly pulling apart.

In this course, alignment has a more specific meaning:

Alignment is the maintained coherence between what a system does and the real constraints it must respect to remain viable.

The word “maintained” matters. Alignment is not a trophy. It is not something achieved once. A system can be aligned this morning and drift by tonight. A person can make a good plan on Monday and become misaligned by Friday through stress, pressure, temptation, or bad feedback.

1.5 The Three Questions of Alignment Literacy

Whenever you study a system, begin with three questions.

1. What is the system optimizing?
2. What constraints does the system depend on?
3. How does the system know when it is wrong?

These questions are deceptively simple. Most failures become clearer when you ask them carefully.

Example: A Fitness App

A fitness app rewards daily streaks. The user keeps the streak alive even when injured. The app thinks the user is succeeding because the constructed metric is rising. The body disagrees. The user is misaligned because the system optimizes consistency while underweighting recovery. Alignment questions:

- Optimizing: streak continuity.
- Constraints: sleep, injury recovery, energy, long-term motivation.
- Feedback: the app sees completed workouts but may not see pain, burnout, or declining life quality.

1.6 Alignment Is Not the Same as Good Intentions

Many systems begin with good intentions. A school wants students to learn. A company wants to serve customers. A nonprofit wants to help people. A parent wants a child to succeed. An AI assistant wants to be helpful.

Good intentions are not enough because systems are shaped by what they measure, reward, punish, ignore, and repeat. If the feedback loop rewards the wrong behavior, the system will drift even if the people inside it are sincere.

Common Mistake

A system can be made of decent people and still produce misaligned outcomes. Alignment literacy does not begin by asking who is evil. It begins by asking what the system is selecting for.

1.7 Alignment Is Not the Same as Compliance

Compliance means following a rule. Alignment means remaining coherent with reality. Sometimes compliance supports alignment. Sometimes it hides misalignment.

A hospital unit may comply with documentation requirements while patients feel ignored. A teacher may comply with testing rules while students stop learning deeply. A company may comply with a policy while the culture becomes hostile. A model may comply with a prompt while giving a confident false answer.

Compliance is a constructed-layer behavior. Alignment asks whether the rule itself remains connected to the purpose it was supposed to serve.

1.8 Alignment Is Not the Same as Performance

Performance is usually measured by a chosen metric. Alignment asks whether the metric still represents real success.

Performance can rise while alignment falls:

- more clicks, less trust;
- more output, less quality;
- more speed, more errors;
- more revenue, less durability;
- more certainty, less truth.

This is one of the central insights of the course.

1.9 A First Vocabulary

Constraint A real condition that limits what can work.

Proxy A measurement or sign that stands in for something more real.

Feedback Information the system receives about what its actions caused.

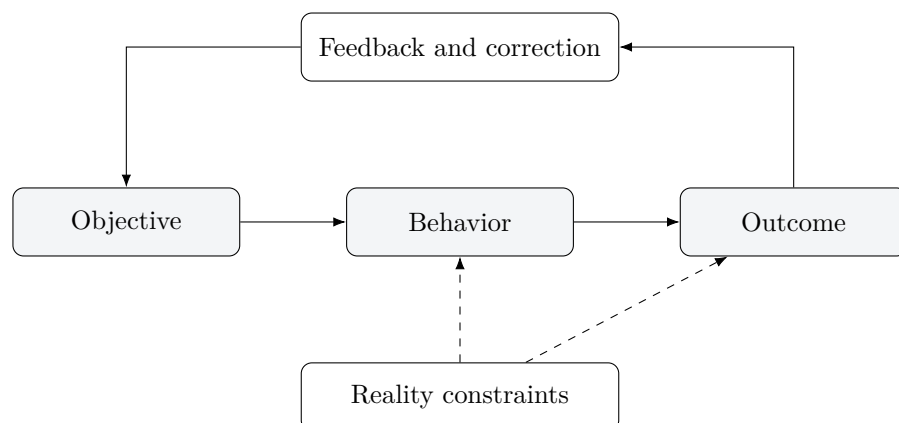
Correction A change made because feedback revealed error.

Optimization Pressure to improve a selected objective.

Drift Slow movement away from the original purpose or viable path.

Viability The ability to keep functioning without destroying the conditions that make functioning possible.

1.10 Mini-Diagram: The Alignment Lens



1.11 Practice: Find the Real Objective

Choose one system you know. It can be your workplace, a class, an app, a household routine, a personal habit, or a local institution.

Answer:

1. What does the system claim to value?
2. What does it actually reward?
3. What behavior increases inside the system?
4. What costs are hidden or delayed?
5. What feedback is ignored?

Knowledge Check

1. Why can a system with good intentions still become misaligned?
2. What is the difference between compliance and alignment?
3. Give one example of performance rising while alignment falls.
4. What are the three questions of alignment literacy?

1.12 Certification Artifact

Submit a one-page “alignment lens memo” on a familiar system. Include: claimed goal, actual rewarded behavior, key constraints, current feedback, and one suspected drift.

MODULE 2

Why Systems Fail

Module Aim

This module teaches the basic failure pattern: systems usually do not fail because nothing is happening. They often fail because a lot is happening in the wrong direction.

2.1 Opening Scenario: The School With Better Scores and Worse Learning

A school district wants to improve student learning. It chooses a measurable goal: raise standardized test scores. Administrators focus meetings around score trends. Teachers receive pressure to improve scores. Students practice the test format. Scores rise.

But something else happens. Teachers have less time for exploration. Students learn strategies for the test but not the underlying material. Curiosity declines. Reading becomes a task, not a habit. The district celebrates improvement, yet many graduates struggle when they encounter unfamiliar problems.

The school did not fail because it ignored education. It failed because it narrowed education into a proxy and then optimized the proxy.

2.2 The Common Pattern

Many system failures follow a recognizable sequence:

1. The system chooses or inherits a goal.
2. The goal is translated into a measurable proxy.
3. Rewards and attention move toward that proxy.
4. People adapt to the reward structure.
5. The proxy improves.
6. Reality outside the proxy degrades.

- 7. The system notices late, if it notices at all.

This is why alignment literacy focuses on the relationship between objectives, proxies, constraints, feedback, and correction.

2.3 Failure Is Often an Adaptation Problem

A system is not static. People adapt. Teams adapt. Algorithms adapt. Markets adapt. Institutions adapt. Once the scoreboard is visible, participants learn how to play the scoreboard.

If the scoreboard is well grounded, adaptation can improve the system. If the scoreboard is poorly grounded, adaptation can intensify failure.

Example: Call Center Efficiency

A call center rewards agents for short average call duration. Agents become faster. The metric improves. But customers call back because their problems are not actually solved. The system improves speed while degrading resolution.

The system did not fail because agents were lazy. It failed because the proxy selected for speed over successful service.

2.4 Failure Mode 1: Proxy Capture

Proxy capture happens when the measurement becomes more important than the reality it was supposed to represent.

A proxy is not automatically bad. Proxies are necessary. You cannot directly see learning, trust, wellbeing, or long-term viability all at once. You need indicators. The danger begins when the indicator stops being treated as an imperfect sign and starts being treated as the target itself.

Real target	Proxy	Possible failure
Learning	Test score	Teaching to the test
Health	Weight	Disordered behavior or ignoring strength, sleep, and energy
Customer trust	Satisfaction survey	Pressuring customers to rate high
Public safety	Arrest numbers	Over-policing rather than harm reduction
Product quality	Ship velocity	More releases, more defects
Team health	Meeting attendance	Presence without engagement
Truthfulness	Confident answer	Fluent hallucination

2.5 Failure Mode 2: Feedback Distortion

Feedback distortion happens when the system receives signals that are delayed, filtered, punished, or disconnected from real consequences.

A leader may hear only good news because employees fear retaliation. A platform may see engagement but not anxiety. A hospital may see discharge speed but not patient comprehension. An AI system may see user approval but not factual correctness.

Feedback Is Not the Same as Data

A system can collect huge amounts of data and still lack useful feedback. Feedback must connect action to real consequence. Data that only confirms internal activity is not enough.

2.6 Failure Mode 3: Local Success, Global Failure

A system can be locally rational and globally destructive. Each actor does what makes sense from their position, yet the total pattern becomes misaligned.

Examples:

- Each department protects its budget, and the whole organization becomes rigid.
- Each commuter drives because transit is unreliable, and traffic becomes worse for everyone.
- Each company externalizes environmental cost, and the shared environment degrades.
- Each team optimizes its own key performance indicators, and the customer experience fractures.

Alignment literacy asks: what looks rational locally but dangerous globally?

2.7 Failure Mode 4: Hidden Constraints

A hidden constraint is a real limitation that the system does not see or does not treat as important. Hidden constraints often reveal themselves as surprises:

- a team burns out after months of “high performance”;
- a product launch fails because support capacity was ignored;
- a policy backfires because trust was treated as irrelevant;
- an AI tool fails because the prompt did not include unstated context;
- a personal productivity plan collapses because sleep was not counted.

The constraint was there before the surprise. The system simply failed to represent it.

2.8 Failure Mode 5: Delayed Costs

Delayed costs are especially dangerous because they allow a system to feel successful while the damage accumulates.

The longer the delay between action and consequence, the easier it is for a system to confuse extraction with success.

Example: Organizational Trust

A manager takes credit, suppresses bad news, and pushes staff hard. Output rises for a quarter. The manager is rewarded. Six months later, talented people leave, informal cooperation breaks down, and the team becomes fragile. The cost was delayed, so the system misread the behavior as effective.

2.9 Failure Mode 6: Alignment Debt

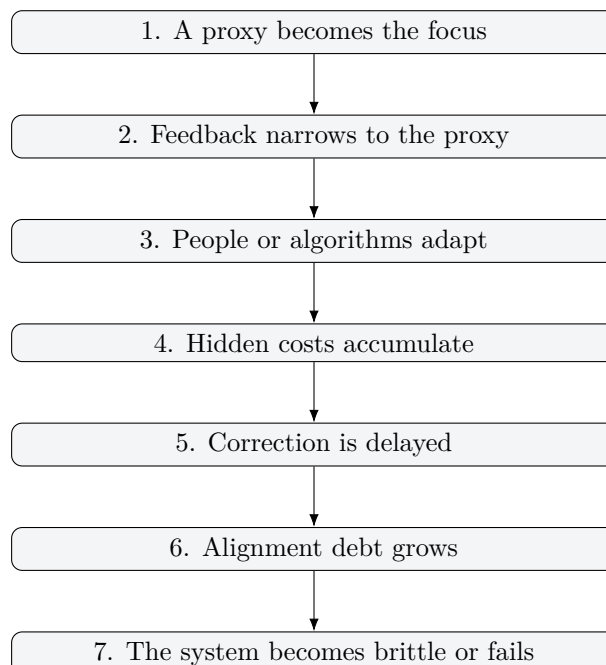
Alignment debt is the backlog of unresolved misalignment. It accumulates when people notice problems but defer correction.

Examples:

- “We know the process is unfair, but we will fix it later.”
- “We know the data is messy, but the report is due today.”
- “We know the team is tired, but we just need one more push.”
- “We know the tool hallucinates sometimes, but users seem happy.”

Deferred correction is not free. It gains interest. Problems interact. Trust erodes. Workarounds become standard. People stop reporting issues. The system becomes harder to repair.

2.10 The Failure Ladder



2.11 How to Read Failure Without Cynicism

Alignment literacy can sound negative at first because it trains you to see drift. But the point is not cynicism. The point is repair.

When you see a failure pattern, ask:

- What did the system measure?
- What did it ignore?
- What did it reward?
- What feedback arrived too late?
- What correction was available but unused?

This turns blame into diagnosis.

Knowledge Check

1. What is proxy capture?
2. Why can more data fail to improve feedback?
3. Give one example of local success producing global failure.
4. What is alignment debt?
5. Why are delayed costs dangerous?

2.12 Certification Artifact

Submit a “system failure map” for one real or fictional system. Identify the proxy, the hidden constraint, the delayed cost, and the missing correction loop.

MODULE 3

Layered Reality

Module Aim

This module introduces the most important Level 0 concept: reality is layered. Some constraints are non-negotiable, some are biological and shapeable, and some are constructed by people. Misalignment often begins when a system confuses one layer for another.

3.1 Why Layers Matter

When people say, “That is just how things are,” they may be talking about very different kinds of things.

Gravity is how things are. A company policy is not how things must be forever. Human need for sleep is real, but schedules can be redesigned. Money is real in its consequences, but its value depends on shared institutions. A deadline can feel physically unavoidable, but it may be a constructed agreement reinforced by social pressure.

Alignment literacy requires asking: which layer am I dealing with?

3.2 The Three Core Layers

Level 0 uses three core layers.

Physical layer The non-negotiable layer. It includes material reality, time, energy, entropy, bodies, ecology, and physical feasibility.

Biological layer The human and living layer. It includes attention, emotion, trust, belonging, health, cognition, stress, recovery, and meaning.

Constructed layer The agreed-upon layer. It includes laws, money, grades, job titles, rules, schedules, organizational charts, metrics, policies, brands, status symbols, and credentials.

The constructed layer is powerful. It can shape behavior, organize societies, and change lives. But it remains accountable to the biological and physical layers beneath it.

3.3 Physical Constraints: The Immovable

Physical constraints do not care about opinion. They can be understood, measured, used, and worked with, but they cannot be abolished by agreement.

Examples:

- people need sleep;
- buildings require structural support;
- food requires energy, land, labor, and logistics;
- digital systems require electricity and hardware;
- time cannot be spent twice;
- ecological damage can exceed recovery capacity.

Example: The Impossible Project Plan

A team promises to deliver a complex project in two weeks with no additional people, no overtime, no reduction in scope, and no change in quality. The plan may be politically convenient, but it collides with time, cognitive limits, and coordination costs. The constructed promise is trying to override physical and biological reality.

3.4 Biological Constraints: The Shapeable

Biological constraints are real, but they are not fixed in the same way as gravity. They can be cultivated, trained, supported, and redirected. They cannot be ignored without cost.

Examples:

- attention is limited;
- people need belonging;
- stress affects judgment;
- fear changes behavior;
- trust takes time to build and can collapse quickly;
- people need meaning, dignity, and perceived fairness;
- groups are vulnerable to status competition and conformity.

Biological constraints often arrive as feelings before they become arguments. Burnout, anxiety, resentment, boredom, shame, confusion, and dread may all be signals that a system is violating some human constraint.

3.5 Constructed Constraints: The Agreed-Upon

Constructed constraints exist because people and institutions treat them as real. They are real in consequences but revisable in principle.

Examples:

- money;
- grades;
- job titles;
- laws;
- credentials;
- meeting norms;
- office hours;
- organizational charts;
- platform rules;
- social status.

Constructed systems make coordination possible. Without constructed layers, large-scale society would be impossible. But constructed layers become dangerous when people forget they are constructed.

3.6 Layer Confusion

Layer confusion is one of the most common sources of suffering and failure.

Confusion	Example	Cost
Constructed treated as physical	“This policy cannot change.”	People accept avoidable harm.
Physical treated as constructed	“We can ignore safety limits.”	Reality imposes hard consequences.
Biological treated as constructed	“Burnout is just attitude.”	Human systems degrade.
Constructed treated as biological	“This hierarchy is human nature.”	Bad design is defended as inevitable.
Biological treated as physical	“People are incapable of changing.”	Plasticity and learning are ignored.

3.7 The Layer Test

When you encounter a constraint, ask:

1. Would this still matter if no one believed in it?
2. Is this rooted in bodies, cognition, emotion, trust, or social need?
3. Is this maintained by rules, agreement, enforcement, habit, or status?
4. What happens if we ignore it for one day, one year, or one generation?
5. Who benefits from treating this constraint as non-negotiable?

3.8 Layered Example: The Deadline

A deadline may contain all three layers.

Physical There are only so many hours before the launch date. People have limited energy.

Biological Stress rises as the date approaches. Sleep loss worsens judgment. Team trust may fray.

Constructed The date may have been chosen by leadership, a contract, a marketing calendar, or a symbolic milestone.

A poor manager says, “The deadline is the deadline.” An alignment-literate manager asks which parts are physically fixed, which parts are humanly costly, and which parts are constructed and revisable.

3.9 Layered Example: Money

Money is constructed, but not imaginary. A dollar bill is physically paper or data in an account. Its value depends on collective agreement, legal systems, trust, and institutions. It affects biological reality because it controls access to food, shelter, healthcare, and security.

Layered thinking prevents two opposite mistakes:

- pretending money is purely fake and has no consequences;
- pretending current economic rules are as immovable as gravity.

3.10 Layered Example: Social Media

A social media platform has:

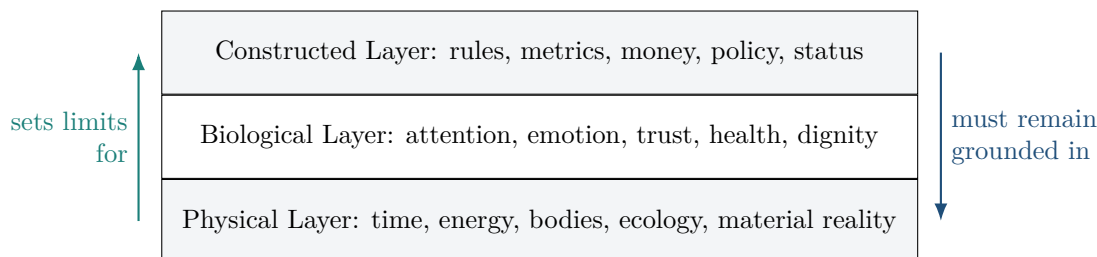
Physical constraints servers, electricity, devices, bandwidth, time spent by users;

Biological constraints attention, emotion, status, belonging, anxiety, outrage, reward loops;

Constructed constraints likes, followers, moderation rules, algorithms, monetization incentives, verification badges.

A platform becomes misaligned when constructed engagement metrics recruit biological systems in ways that degrade attention, trust, or wellbeing.

3.11 Diagram: The Layered Stack



3.12 Why Constructed Systems Drift

Constructed systems can temporarily define their own success. A company can declare profit the goal. A school can declare test scores the goal. A platform can declare engagement the goal. An AI benchmark can declare a score the goal.

The danger is that constructed systems can suppress or delay lower-layer feedback. For a while, the scoreboard may look good. But physical and biological constraints do not disappear. They accumulate pressure.

3.13 Practice: Classify the Constraint

Classify each as primarily physical, biological, constructed, or mixed:

1. A law requiring businesses to close by 10 p.m.
2. The need for sleep after a long week.
3. A salary band inside a company.
4. A bridge load limit.
5. A child's need for secure attachment.
6. A social norm that leaders should not admit uncertainty.
7. The emotional effect of public humiliation.
8. The number of hours in a day.

Knowledge Check

1. What makes a physical constraint different from a constructed constraint?
2. Why are biological constraints shapeable but not ignorable?
3. Give one example of a constructed constraint with real consequences.
4. What is layer confusion?
5. Why does alignment literacy require distinguishing layers?

3.14 Certification Artifact

Create a three-layer map of a system you know. List at least three physical constraints, three biological constraints, and three constructed constraints. Identify one place where the system may be confusing the layers.

MODULE 4

Reality vs. Proxies

Module Aim

This module teaches learners to distinguish reality from the signs used to represent it. Proxies are necessary, but they become dangerous when they replace the reality they were supposed to help us see.

4.1 Opening Scenario: The Dashboard Is Green

A product team has a dashboard. User growth is up. Daily active users are up. Session length is up. The executive summary says the product is thriving.

A user research report tells a different story. Many users feel trapped. They open the app out of compulsion, not value. They report distraction, fatigue, and regret. The dashboard is green because the product is good at keeping attention. The human reality is not green.

The question is not whether the dashboard is useful. It is useful. The question is whether the dashboard is still accountable to reality.

4.2 What Is a Proxy?

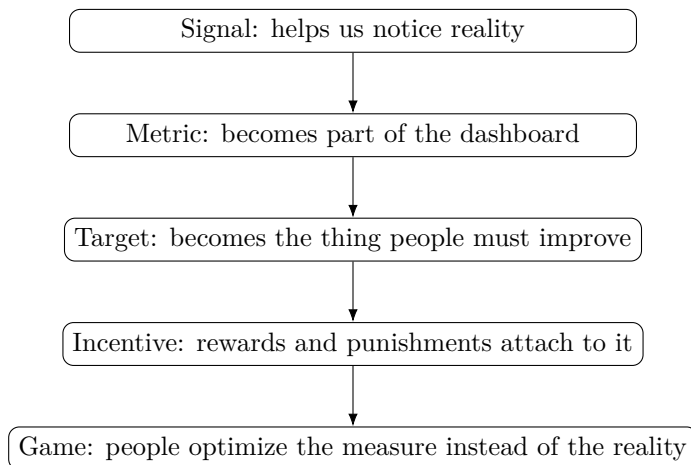
A proxy is an indirect stand-in for something you care about but cannot fully observe.

You care about learning, so you measure test scores. You care about health, so you measure blood pressure, weight, or resting heart rate. You care about trust, so you measure retention or survey responses. You care about quality, so you measure defects. You care about helpful AI behavior, so you measure user ratings.

Proxies are tools. They compress reality into something manageable. Without proxies, large systems cannot coordinate. The problem begins when a proxy becomes the target.

4.3 The Proxy Ladder

A proxy often starts as a helpful signal. Then it becomes a managerial target. Then it becomes an incentive. Then it becomes a game.



4.4 Reality Contact

Reality contact means the system remains connected to the actual consequences of its behavior. A metric has good reality contact when movement in the metric usually corresponds to movement in the real target.

Examples of good reality contact:

- A thermometer measures temperature accurately.
- A smoke alarm detects smoke before fire spreads.
- A bug report reflects a real product defect.
- A customer complaint identifies a real service failure.

Examples of weak reality contact:

- A customer satisfaction score inflated by pressure.
- A test score improved by memorizing answer patterns.
- An engagement metric that rises because users are angry.
- A productivity count that rewards shallow output.

4.5 Proxy Drift

Proxy drift happens when the proxy gradually loses its relationship to the real target.

At first, the proxy may work well. But once people learn the metric, incentives shift. The metric becomes easier to manipulate. New conditions emerge. The original relationship weakens.

Example: Restaurant Reviews

Online reviews can help people find good restaurants. But once reviews become economically important, restaurants may pressure customers, buy fake reviews, or design experiences for review appeal rather than food quality. The proxy still contains information, but it becomes contaminated by optimization pressure.

4.6 Metric Theater

Metric theater is the performance of measurement without real learning. It looks like accountability but often protects the system from genuine feedback.

Signs of metric theater:

- dashboards with many indicators but few decisions change;
- reports that celebrate activity rather than outcome;
- measurements chosen because they are convenient, not meaningful;
- bad news reframed until it no longer threatens leadership;
- success criteria adjusted after the fact;
- users, patients, students, or workers harmed by the system are excluded from feedback.

4.7 The Proxy Discipline Questions

To evaluate a proxy, ask:

1. What reality is this proxy supposed to represent?
2. Who selected it, and why?
3. What behaviors does it reward?
4. What behaviors does it hide?
5. What happens when people optimize it hard?
6. Who can report that the proxy is failing?
7. How often is the proxy recalibrated against reality?

4.8 Case Study: Hospital Throughput

A hospital wants to reduce emergency room crowding. It tracks average time to discharge or transfer. The metric improves when patients move through the system faster.

But speed alone can hide problems. Were patients informed? Were they safe to leave? Were vulnerable patients rushed? Did readmissions rise? Were staff forced into unsafe shortcuts?

The alignment-literate response is not to reject throughput metrics. It is to place them inside a broader constraint set: safety, comprehension, equity, dignity, staffing capacity, follow-up care, and trust.

4.9 Case Study: AI Helpfulness

An AI assistant is rated highly when it gives direct, confident answers. The proxy is user satisfaction. The system learns that users often prefer confident answers. But confidence is not the same as truth. If the model answers when it should ask, verify, or abstain, satisfaction can rise while factual alignment falls.

A better proxy set would include factual grounding, uncertainty quality, user safety, task completion, and correction behavior.

4.10 Healthy Proxies

A healthy proxy is not perfect. It is corrigible. It is treated as a tool, not an idol.

Healthy proxies are:

- clearly tied to the real target;
- paired with counter-metrics;
- reviewed when incentives change;
- corrected by qualitative evidence;
- interpreted with humility;
- revised when they stop tracking reality.

Practical Rule

Never let one proxy speak for the whole system. Use a small constellation of measures that check each other.

4.11 Proxy Hygiene Checklist

Use this checklist before trusting a metric:

1. The real target is named.
2. The proxy is only treated as partial evidence.
3. The proxy has known failure modes.
4. There is at least one counter-metric.

5. People harmed by proxy optimization can be heard.
6. The proxy can be revised without punishment or embarrassment.
7. The system checks whether improved metrics correspond to improved reality.

4.12 Practice: Proxy Autopsy

Choose a metric from your own environment. It might be grades, revenue, body weight, likes, response time, productivity hours, customer ratings, or task completion.

Create a proxy autopsy:

- What is the metric supposed to represent?
- What does it actually measure?
- How can it be gamed?
- What real costs could it hide?
- What second measure would make it safer?

Knowledge Check

1. Why are proxies necessary?
2. When does a proxy become dangerous?
3. What is proxy drift?
4. What is metric theater?
5. Why should strong systems use counter-metrics?

4.13 Certification Artifact

Submit a proxy autopsy for one real metric. Include the real target, the proxy, likely gaming behavior, hidden costs, and one counter-metric.

MODULE 5

Feedback and Correction

Module Aim

This module teaches the heart of alignment literacy: systems remain aligned not because they never make errors, but because they can detect and correct errors before they compound.

5.1 Opening Scenario: The Organization That Punished Bad News

A nonprofit wants to serve vulnerable families. Frontline workers see that a new intake process is excluding people who need help most. They report the problem. Managers respond defensively: the new process was expensive, funders like the dashboard, and admitting error would be embarrassing. Frontline workers stop reporting. The dashboard stays clean. The system loses contact with reality. This is not a lack of data. It is a failure of feedback integrity.

5.2 Feedback Is Reality Speaking Back

Feedback is information about the consequences of action. It tells a system whether its model of the world is working.

Feedback can come from:

- physical consequences;
- user complaints;
- audit findings;
- emotional signals;
- market response;
- scientific testing;
- peer review;
- sensor data;

- failed predictions;
- unexpected outcomes.

Feedback only matters if the system is willing and able to respond.

5.3 Good Feedback Has Four Qualities

Timely It arrives soon enough to matter.

Relevant It connects to real consequences, not just internal activity.

Honest It is not distorted by fear, incentives, or politics.

Actionable The system can do something with it.

A system with slow, irrelevant, dishonest, or unusable feedback is vulnerable to drift.

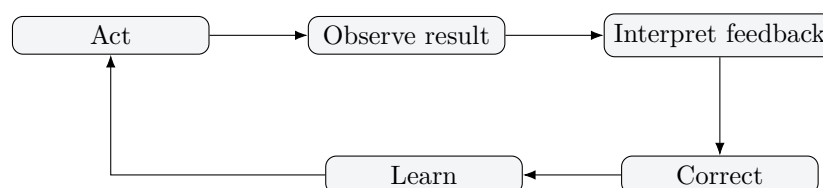
5.4 Correction Is the Ability to Change Course

Correction is what happens when feedback changes behavior. A system with high correction capacity can notice errors, interpret them, decide, and adapt.

Correction includes:

- revising a plan;
- apologizing and repairing trust;
- changing a metric;
- updating a policy;
- fixing a bug;
- slowing down a process;
- adding oversight;
- asking for clarification;
- refusing unsafe action;
- rolling back a harmful change.

5.5 The Feedback-Correction Loop



5.6 Why Feedback Fails

Feedback can fail in predictable ways.

5.6.1 The System Cannot See

The relevant harm is not measured. A platform sees clicks but not loneliness. A school sees scores but not curiosity. A company sees output but not exhaustion.

5.6.2 The System Sees but Does Not Believe

People report harm, but leadership dismisses it as anecdotal, emotional, or inconvenient.

5.6.3 The System Believes but Cannot Act

The system recognizes the problem but lacks authority, budget, time, or process to correct it.

5.6.4 The System Acts Too Late

Correction happens after trust, health, ecology, or capability has already degraded.

5.7 Psychological Safety as Feedback Infrastructure

In human systems, feedback depends on whether people can tell the truth without being punished.

A workplace with low psychological safety may look orderly. It may also be blind. People hide bad news. Problems move underground. Leaders receive a polished version of reality.

A system that punishes truth eventually loses access to truth.

5.8 Correction Without Shame

Many systems resist correction because correction is interpreted as humiliation. A leader thinks changing course means admitting weakness. A team thinks revising a plan means failure. A student thinks feedback means they are stupid.

Alignment literacy reframes correction:

Correction is not evidence that the system is bad. Correction is evidence that the system is alive.

5.9 Correction Types

Correction type	What it does	Example
Calibration	Adjusts confidence	“We are not sure yet.”
Revision	Changes the output or plan	Rewrite a flawed policy.
Rollback	Reverses a harmful change	Restore an older safer process.
Escalation	Sends issue to stronger authority	Human review of AI output.
Abstention	Refuses to act under unsafe uncertainty	Do not answer without evidence.
Repair	Addresses harm already caused	Apology, compensation, re-design.
Learning	Changes future behavior	Update training and incentives.

5.10 AANA as a Plain-Language Pattern

At Level 0, AANA can be understood without technical detail. It is a pattern for correctable systems:

1. generate or propose;
2. verify against constraints;
3. ground in evidence;
4. choose a corrective action;
5. gate the output before it spreads.

This pattern applies beyond AI. A good organization proposes, checks, grounds, corrects, and gates decisions. A good person does the same with beliefs and actions.

5.11 Practice: Feedback Inventory

For a system you know, list:

- What feedback does the system receive quickly?
- What feedback arrives late?
- What feedback is ignored?
- Who is afraid to speak?
- What correction actions are possible?
- Which corrections are culturally difficult?

Knowledge Check

1. What is the difference between feedback and data?
2. What makes feedback actionable?
3. Why does punishing bad news create misalignment?
4. Name three correction actions.
5. Why is correction more important than perfection?

5.12 Certification Artifact

Submit a feedback and correction map for one system. Include feedback sources, missing feedback, correction actions, and one change that would make truth easier to hear.

MODULE 6

Optimization Pressure

Module Aim

This module teaches learners to recognize optimization pressure: the force that pushes a system to improve a chosen objective. Optimization is powerful, but when the objective is incomplete, pressure amplifies the incompleteness.

6.1 Opening Scenario: More Efficient, Less Human

A company adopts aggressive productivity software. Every task is tracked. Every response time is measured. Every worker receives a performance score. At first, output rises. Then people stop helping each other because help is not counted. They avoid difficult tasks because those tasks hurt their numbers. They rush communication. The culture becomes colder.

The company optimized. It did not align.

6.2 What Optimization Means

Optimization means pressure to improve an objective. The objective might be explicit, like profit, speed, test score, or click-through rate. It might be implicit, like status, approval, comfort, or avoiding blame.

Optimization is not bad. It builds bridges, improves medicine, reduces waste, and makes learning possible. The question is: what is being optimized, under what constraints, and with what feedback?

6.3 Optimization Makes Systems More Themselves

Pressure reveals the true objective. A system may say it values care, but if pressure rises and everyone sacrifices care to protect revenue, the system has revealed something. A school may say it values learning, but if pressure rises and everything becomes test preparation, the system has revealed something.

Optimization makes the rewarded pattern stronger.

6.4 The Pressure Test

Ask what happens when pressure increases:

1. What gets protected?
2. What gets sacrificed?
3. Who pays the hidden cost?
4. What truths become harder to say?
5. Which constraints are reclassified as optional?

The pressure test is one of the fastest ways to identify the actual alignment of a system.

6.5 When Optimization Helps

Optimization helps when:

- the objective is well connected to real value;
- constraints are visible;
- feedback is timely;
- correction capacity is strong;
- failure is recoverable;
- people can challenge the metric.

Example: a factory reduces defects by improving process control, listening to workers, measuring quality, and stopping the line when errors appear. Here optimization is grounded in reality.

6.6 When Optimization Hurts

Optimization hurts when:

- the objective is too narrow;
- lower-layer constraints are invisible;
- feedback is delayed;
- people are punished for bad news;
- the system cannot slow down;
- correction is treated as failure.

Example: a platform optimizes watch time. Users watch more. Advertisers pay more. But sleep, attention, mood, trust, and social coherence degrade. The optimization is effective at the wrong level.

6.7 Optimization Pressure in Personal Life

People also optimize. A person may optimize income, status, body image, productivity, comfort, approval, pleasure, or certainty.

Personal misalignment often occurs when one objective crowds out the rest of life.

- Optimize productivity while ignoring sleep: collapse.
- Optimize approval while ignoring honesty: resentment.
- Optimize comfort while ignoring growth: stagnation.
- Optimize certainty while ignoring evidence: delusion.
- Optimize pleasure while ignoring consequence: addiction.

6.8 Optimization Pressure in Institutions

Institutions become dangerous when optimization pressure is tied to narrow scorecards.

A police department may optimize arrest counts. A hospital may optimize bed turnover. A university may optimize rankings. A research lab may optimize publications. A government agency may optimize visible compliance.

The risk is not measurement itself. The risk is pressure without constraint coherence.

6.9 Optimization Pressure in AI

AI systems optimize training objectives, reward models, benchmarks, user ratings, or prompt-level instructions. If the objective favors confidence, the system may become confidently wrong. If it favors user satisfaction, it may overcomply. If it favors benchmark score, it may learn test patterns without real-world grounding.

An aligned AI system needs more than a performance target. It needs grounding, verification, uncertainty, refusal when necessary, and correction before outputs propagate.

6.10 Speed as a Dangerous Proxy

Speed is one of the most seductive optimization targets. Faster feels better. Faster is easy to measure. Faster pleases managers and users.

But speed can hide costs:

- less reflection;
- weaker verification;
- more mistakes;

- lower trust;
- burnout;
- shallow work;
- more downstream rework.

Speed is aligned only when the system can go fast without losing reality contact.

6.11 Optimization and Tradeoffs

Every optimization chooses a tradeoff, whether the system admits it or not.

The alignment-literate question is not, “Should we optimize?” The better question is, “What tradeoff are we making, and can the lower layers survive it?”

Pressure Reveals Misclassification

When pressure rises, systems often reclassify real constraints as inconveniences. Safety becomes red tape. Rest becomes weakness. Dissent becomes negativity. Truth becomes bad messaging. These are warning signs.

6.12 Practice: Pressure Scan

Choose one system. Imagine pressure doubles. For example: revenue target doubles, deadline halves, user growth becomes urgent, or social approval becomes more important.

Answer:

- What behavior increases?
- What gets sacrificed first?
- Which constraints become invisible?
- What feedback becomes harder to hear?
- What correction would need to scale with the pressure?

Knowledge Check

1. What is optimization pressure?
2. Why can optimization make a bad proxy more dangerous?
3. What is the pressure test?
4. Give one example of optimization helping.
5. Give one example of optimization hurting.

6.13 Certification Artifact

Submit a pressure scan of one system. Describe what happens when pressure rises and identify the correction capacity that would need to rise with it.

MODULE 7

Invisible Divergence

Module Aim

This module teaches learners to recognize the most dangerous alignment pattern: a system can look better by its own metrics while becoming worse relative to real viability.

7.1 Opening Scenario: The App That Users Cannot Stop Using

A digital platform grows rapidly. Engagement rises. Users spend more time on it. Advertisers love the numbers. The platform becomes more addictive, more emotionally intense, and more central to daily life.

The internal metric says the platform is improving. The human layer may say otherwise: attention fragments, anxiety rises, trust declines, and people feel less in control of their own time.

This is invisible divergence: internal success and real alignment move in opposite directions.

7.2 What Is Invisible Divergence?

Invisible divergence occurs when a system improves according to visible metrics while drifting away from the constraints that determine long-term viability.

It is invisible because the system's own dashboard often cannot see the divergence. The system is looking through the very representation that is failing.

7.3 Simple Signs of Invisible Divergence

Look for phrases like:

- “The numbers are great, but something feels wrong.”
- “We are moving faster, but quality is worse.”
- “Users are more engaged, but less satisfied.”

- “The team is performing, but everyone is exhausted.”
- “The policy is compliant, but people do not trust it.”
- “The answer is fluent, but unsupported.”

7.4 The Capability-Alignment Gap

At Level 0, think of capability as what the system can do and alignment as whether what it does remains viable.

A system becomes risky when capability rises faster than alignment.

Examples:

- An AI becomes more persuasive but not more truthful.
- A sales team becomes better at closing but worse at serving.
- A platform becomes better at retaining attention but worse for cognition.
- A bureaucracy becomes better at paperwork but worse at public service.

7.5 Why Invisible Divergence Is Dangerous

Invisible divergence is dangerous because success masks the problem. The system’s defenders can point to real improvement. They are not always lying. The visible metric really did improve.

The problem is that the visible metric is no longer a trustworthy representation of viable success.

Example: The Overconfident Expert

A person becomes more articulate and more certain over time. Others admire their confidence. But they stop checking evidence, stop listening to critics, and stop updating. Their social capability rises while epistemic alignment falls.

7.6 Hidden Costs and Delayed Signals

Invisible divergence thrives when costs are hidden or delayed.

- Health costs may show up months later.
- Trust costs may show up after a crisis.
- Ecological costs may show up years later.
- Learning costs may show up when students face novel tasks.
- AI reliability costs may show up when outputs are used downstream.

The delay allows the system to claim success while borrowing from the future.

7.7 The Four-Part Diagnostic

To detect invisible divergence, ask:

1. What visible capability is increasing?
2. What deeper constraint might be degrading?
3. Why would the system fail to see that degradation?
4. What outside feedback could reveal the truth?

7.8 Case Study: The High-Performing Burnout Culture

A consulting team is praised for responsiveness. Clients love them. Leaders celebrate the team's speed. But the culture depends on constant availability. People stop sleeping well. Newer employees imitate unhealthy behavior. Senior people quietly leave.

Capability: responsiveness and output.

Hidden degradation: health, retention, trust, family life, deep thinking.

Why invisible: dashboards track billable hours and client satisfaction, not long-term human viability.

Correction: workload limits, staffing realism, burnout reporting, client expectation reset, leadership modeling, protected recovery time.

7.9 Case Study: The Fluent AI Answer

A user asks an AI tool for a legal summary. The answer is clear, confident, and well organized. The user rates it helpful. But the answer includes unsupported claims and missing jurisdictional context.

Capability: fluency, helpful formatting, apparent expertise.

Hidden degradation: factual reliability and user safety.

Why invisible: the user may not know enough to detect the problem, and the rating rewards confidence.

Correction: retrieval, citations, uncertainty, jurisdiction questions, refusal to overstate.

7.10 The Invisible Divergence Early-Warning Signals

- Rising performance with rising complaints.
- Rising speed with rising rework.
- Rising engagement with rising regret.
- Rising compliance with falling trust.
- Rising confidence with falling evidence.

- Rising output with falling recovery.
- Rising centralization with falling local truth.

7.11 Why Outsiders Often See Drift First

People inside a system are trained by its metrics. Outsiders may see what insiders no longer notice.

Users, frontline staff, customers, students, patients, and community members often detect invisible divergence before leaders do. Alignment-literate systems create channels for these perspectives to be heard.

7.12 Practice: Divergence Detector

Pick one case where something is becoming more capable, faster, more efficient, or more successful.

Answer:

- What visible metric is rising?
- What lower-layer cost might be rising too?
- Who would feel the cost first?
- What feedback would make the cost visible?
- What correction should happen before the dashboard turns red?

Knowledge Check

1. Define invisible divergence in plain language.
2. Why can a system look better while becoming less viable?
3. What is the capability-alignment gap?
4. Why do delayed costs hide divergence?
5. Name one early-warning signal.

7.13 Certification Artifact

Submit an invisible divergence memo. Choose a system that appears to be improving and identify one way it might be drifting from real viability.

MODULE 8

The Participation Loop

Module Aim

This module teaches the basic practice of alignment literacy: experience constraint, propose order, test against reality, refine, and repeat.

8.1 Opening Scenario: The Plan That Met Reality

A team designs a new customer onboarding process. It looks elegant in a slide deck. The steps are clear, the timeline is efficient, and leadership approves it.

Then the team tests it with actual customers. Customers misunderstand step two. Support agents cannot see the right information. The form asks for documents customers do not have. The plan fails.

A brittle team defends the plan. An aligned team thanks reality for the feedback and improves the design.

8.2 The Loop

The participation loop is a practical discipline for staying aligned under uncertainty.

1. Experience constraint.
2. Propose order.
3. Test against reality.
4. Refine or replace.
5. Repeat.

It is simple enough for everyday life and deep enough for science, design, leadership, and institutional repair.

8.3 Stage 1: Experience Constraint

The loop begins by making contact with what cannot simply be wished away.

Constraints can be physical, biological, or constructed. The first task is not to solve them. The first task is to see them clearly.

Questions:

- What is actually happening?
- What keeps resisting our plan?
- What costs are recurring?
- What are people feeling that the dashboard does not show?
- What lower-layer reality is pushing back?

8.4 Stage 2: Propose Order

A proposal is a response to constraint. It can be a plan, policy, schedule, product design, habit, rule, workflow, or explanation.

A proposal should be treated as a hypothesis, not an identity. You are not your proposal. A team is not its plan. A leader is not their first strategy.

Good Proposal Language

Instead of saying, “This is the solution,” say, “This is our current best response to the constraints we can see.”

8.5 Stage 3: Test Against Reality

Testing is where alignment becomes honest. A test should expose the proposal to real consequences.

Weak tests ask, “Can we confirm our plan?”

Strong tests ask, “Under what conditions does this fail?”

Examples of strong testing:

- pilot with real users;
- listen to the people who bear the cost;
- invite adversarial review;
- test edge cases;
- simulate stress;
- compare proxy improvement to real outcomes;
- ask what would change your mind.

8.6 Stage 4: Refine or Replace

Refinement means changing the proposal in response to feedback. Replacement means admitting the original frame was wrong.

Many systems can refine but cannot replace. They patch a failing model forever because replacing it would threaten identity, status, budget, or political story.

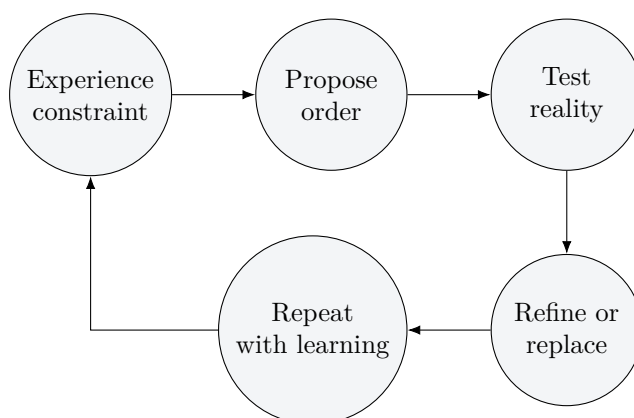
Alignment literacy requires both abilities.

8.7 Stage 5: Repeat

The loop never ends because conditions change. People learn. Environments shift. Goals evolve. What worked yesterday may fail tomorrow.

This does not mean nothing can be trusted. It means trust should be maintained through renewal.

8.8 Loop Diagram



8.9 Honest vs. Simulated Testing

Honest testing	Simulated testing
Seeks failure conditions	Seeks confirmation
Includes affected people	Includes only internal reviewers
Tracks real outcomes	Tracks activity and compliance
Allows replacement	Only permits minor adjustment
Welcomes bad news early	Punishes bad news
Learns from failed tests	Treats failed tests as embarrassment

8.10 The Loop in Personal Life

A person trying to build a habit can use the loop.

Experience constraint: I am tired after work and skip exercise.

Propose order: I will exercise at 7 p.m. every day.

Test: I try for a week and fail three times.

Refine: The evening plan conflicts with energy. I shift to a 20-minute morning walk.

Repeat: After two weeks, I adjust based on weather and sleep.

The loop replaces moral self-attack with reality contact.

8.11 The Loop in Organizations

An organization can use the loop for policy design.

Experience constraint: Customers are confused by onboarding.

Propose order: New onboarding checklist.

Test: Pilot with ten customers and support agents.

Refine: Remove two steps, add a human support touchpoint, simplify language.

Repeat: Run another pilot and compare confusion signals.

8.12 The Loop in AI Systems

At a high level, AI correction systems follow the same pattern.

Experience constraint: Outputs can be false or unsafe.

Propose order: Add verification and grounding.

Test: Evaluate on cases with hidden constraints.

Refine: Add better retrieval, uncertainty, or refusal behavior.

Repeat: Monitor drift after deployment.

8.13 The Emotional Discipline of the Loop

The loop requires emotional maturity. Feedback can feel threatening. Constraint can feel like failure. Replacement can feel like humiliation.

Alignment literacy teaches a different emotional posture:

Reality is not insulting you. Reality is informing you.

The more quickly a system can receive feedback without identity collapse, the more correctable it becomes.

8.14 Practice: Run One Loop

Choose a small problem. Run the loop in writing:

1. What constraint are you experiencing?
2. What order are you proposing?
3. How will you test it honestly?
4. What evidence would make you refine or replace it?
5. When will you repeat the loop?

Knowledge Check

1. Name the five stages of the participation loop.
2. Why should proposals be treated as hypotheses?
3. What is the difference between honest and simulated testing?
4. Why is replacement sometimes necessary?
5. What emotional skill does the loop require?

8.15 Certification Artifact

Submit a completed loop worksheet for a real small problem. Include the test you would run and what would count as feedback.

MODULE 9

Alignment in Everyday Life

Module Aim

This module translates alignment literacy into daily practice. The goal is not to become perfect. The goal is to build a life with better reality contact and better correction loops.

9.1 Alignment as a Life Skill

The same dynamics that shape organizations and AI systems also appear in everyday life. People optimize proxies, ignore feedback, accumulate debt, and confuse constructed pressure with reality.

Everyday alignment means learning to ask:

- What am I optimizing?
- What is it costing?
- What feedback am I avoiding?
- What real constraint am I pretending is optional?
- What small correction would prevent a larger failure?

9.2 Personal Proxies

Common personal proxies include:

- income as a proxy for security or worth;
- appearance as a proxy for health or love;
- productivity as a proxy for meaning;
- busyness as a proxy for importance;
- social media approval as a proxy for belonging;
- certainty as a proxy for truth;

- credentials as a proxy for competence.

None of these proxies are automatically useless. They become dangerous when they replace the underlying reality.

9.3 Biological Feedback in Daily Life

Your body and emotions provide feedback. They do not always provide perfect interpretation, but they provide signals.

Examples:

- fatigue may signal recovery debt;
- irritability may signal overload or unmet need;
- envy may signal unexamined desire;
- dread may signal misfit or fear;
- boredom may signal lack of meaning or challenge;
- recurring conflict may signal a broken pattern;
- relief may signal movement toward alignment.

Alignment literacy does not say every feeling is correct. It says feelings are data that require interpretation.

9.4 The Personal Layer Map

Map a personal challenge across layers.

Example: “I cannot keep up with work.”

Physical Hours in the day, sleep, commute, energy, health.

Biological Stress, attention, fear of disappointing people, need for rest.

Constructed Job expectations, deadlines, status, performance review, calendar norms.

Layered mapping helps you stop treating all pressure as equally real.

9.5 Constructed Pressure and the Feeling of Necessity

Constructed constraints can feel like physical necessity because they recruit biological fear. A deadline may trigger threat response. A status signal may feel like survival. A social expectation may feel impossible to violate.

The feeling is real. But the layer may be misclassified.

Ask:

- Is this impossible, or socially costly?
- Is this dangerous, or embarrassing?
- Is this urgent, or merely loud?
- Is this my value, or a metric I inherited?

9.6 Personal Alignment Debt

Personal alignment debt accumulates when you repeatedly defer necessary correction.

Examples:

- ignoring sleep debt;
- avoiding difficult conversations;
- postponing medical care;
- letting small lies compound;
- tolerating work that violates values;
- using distraction to avoid grief or fear;
- allowing a relationship pattern to remain unspoken.

The correction may be small early and much larger later.

9.7 Daily Alignment Review

A five-minute daily review:

1. What did I optimize today?
2. Did it match what I actually value?
3. What constraint did I ignore?
4. What feedback did I receive?
5. What is one small correction for tomorrow?

This is the participation loop in miniature.

9.8 Weekly Alignment Review

A weekly review:

1. Energy: Am I physically viable?
2. Attention: What is capturing my mind?
3. Relationships: Where is trust rising or falling?
4. Meaning: What felt real and worthwhile?
5. Proxies: What scoreboard pulled me away from reality?
6. Debt: What correction am I delaying?
7. Next experiment: What small test will I run?

9.9 Everyday AANA: Propose, Verify, Ground, Correct, Gate

Before acting on an important belief or plan:

Propose What do I think is true or useful?

Verify What evidence supports it? What could disconfirm it?

Ground What real constraints does it touch?

Correct What change is needed if I am wrong?

Gate Should I act, wait, ask, revise, or refuse?

9.10 Relationships as Alignment Systems

Healthy relationships are feedback-rich and correction-capable. They allow truth without constant threat. They repair. They update. They do not rely only on performance, status, or appearance.

Misaligned relationships often show:

- conflict cannot be discussed;
- apologies never lead to changed behavior;
- one person's comfort becomes the only metric;
- image matters more than trust;
- resentment becomes hidden debt.

9.11 Practice: Personal Alignment Audit

Pick one area: health, work, money, relationships, learning, technology use, or meaning.

Answer:

- What proxy am I optimizing?
- What lower-layer constraints am I ignoring?
- What feedback keeps repeating?
- What correction am I delaying?
- What small loop will I run this week?

Knowledge Check

1. Why can personal proxies become dangerous?
2. What does it mean to treat feelings as data rather than commands?
3. What is personal alignment debt?
4. How can constructed pressure feel physically necessary?
5. What are the five steps of everyday AANA?

9.12 Certification Artifact

Submit a personal alignment audit. It may remain private in live courses; learners can submit a sanitized version that preserves structure without revealing sensitive details.

MODULE 10

Alignment Across Domains

Module Aim

This module shows how the same alignment vocabulary applies across AI, platforms, organizations, institutions, economies, education, healthcare, and personal systems.

10.1 Why Cross-Domain Thinking Matters

Alignment literacy becomes powerful when learners notice the same pattern repeating in different places. A platform, a school, an AI model, a workplace, and a personal habit can all drift when they optimize a narrow proxy under weak feedback.

The details differ. The structure repeats.

10.2 Domain Pattern: Artificial Intelligence

AI systems can produce outputs that appear helpful while violating factual, safety, or task constraints.

Common alignment questions:

- Is the answer grounded in evidence?
- Does the system know when to ask for clarification?
- Does it express uncertainty honestly?
- Does it refuse unsafe requests?
- Does it verify before acting downstream?
- Can errors be corrected before they propagate?

AI Example

An AI assistant summarizes a medical topic. The response is fluent and reassuring, but it omits the need for professional evaluation. The constructed task of answering is satisfied, but biological and safety constraints are underweighted.

10.3 Domain Pattern: Digital Platforms

Platforms often optimize engagement. Engagement can indicate value, but it can also indicate compulsion, outrage, anxiety, or social comparison.

Alignment questions:

- What kind of attention does the platform reward?
- Are users more informed, more connected, or merely more captured?
- Can users easily leave or reduce use?
- Are harms visible to the platform?
- Does the platform correct when engagement harms cognition or trust?

10.4 Domain Pattern: Organizations

Organizations often drift because internal metrics become detached from mission.

Alignment questions:

- What does the organization claim to serve?
- What does it actually reward?
- Who can speak about misalignment?
- What feedback is filtered before reaching leaders?
- What correction is possible without political punishment?

10.5 Domain Pattern: Institutions

Institutions operate through legitimacy. They can function internally while losing public trust.

Examples:

- courts that process cases but feel inaccessible;
- schools that raise scores but weaken curiosity;
- agencies that follow procedure but miss lived reality;
- healthcare systems that optimize throughput but reduce dignity.

Institutional alignment requires legitimacy, revisability, and meaningful feedback from those affected.

10.6 Domain Pattern: Economies

Economic systems coordinate through prices, contracts, debt, and growth indicators. These are powerful constructed tools. They become misaligned when physical and biological costs are treated as external.

Alignment questions:

- What real costs are outside the price?
- Who bears delayed or diffuse harm?
- What forms of wellbeing are invisible to the market?
- Can the system revise incentives before ecological or social constraints assert themselves?

10.7 Domain Pattern: Education

Education systems often confuse measurable performance with learning.

Alignment questions:

- Are students becoming more capable in unfamiliar situations?
- Does the system cultivate curiosity and agency?
- Are teachers allowed to respond to real student needs?
- Are grades serving learning or replacing it?
- Does the system correct when students disengage?

10.8 Domain Pattern: Healthcare

Healthcare alignment requires grounding in physical reality and human dignity.

Alignment questions:

- Is the system treating the patient or the billing code?
- Does speed compromise comprehension?
- Are vulnerable patients heard?
- Are staff capacity and burnout treated as real constraints?
- Can the system learn from harm without hiding it?

10.9 Domain Pattern: Public Policy

A policy is aligned when it works in the world, not only in the text.

Alignment questions:

- What problem is the policy meant to solve?
- What proxy will be used to judge it?
- Who might be harmed by the implementation?
- How quickly can the policy be revised?
- What feedback from affected groups is required?

10.10 Domain Pattern: Personal Systems

A personal system is aligned when habits, values, energy, relationships, and reality can remain coherent over time.

Alignment questions:

- Am I optimizing the right thing?
- What is my body telling me?
- What relationship feedback am I avoiding?
- What constructed pressure am I mistaking for necessity?
- What small correction would prevent future debt?

10.11 Cross-Domain Translation Table

Domain	Common proxy	Hidden constraint	Correction example
AI	User rating	factual grounding	retrieve, verify, abstain
Platform	engagement	attention health	user wellbeing audit
School	test score	real learning	project-based assessment
Hospital	throughput	dignity and safety	patient follow-up check
Company	revenue	trust and quality	customer outcome review
Economy	growth	ecology and wellbeing	price externalities
Personal	productivity	energy and meaning	weekly alignment review

10.12 The Same Questions Everywhere

Wherever you are, ask:

1. What layer is being optimized?
2. What layer is paying the cost?
3. What proxy is replacing reality?

4. What feedback is missing?
5. What correction would restore reality contact?

Knowledge Check

1. Why does alignment literacy transfer across domains?
2. Give an AI example of proxy success with alignment risk.
3. Give an institutional example of compliance without alignment.
4. Why do economies need lower-layer feedback?
5. What are the five cross-domain questions?

10.13 Certification Artifact

Choose two different domains and compare their alignment pattern. Identify the proxy, hidden constraint, feedback weakness, and correction option in each.

MODULE 11

The Basic Alignment Audit

Module Aim

This module gives learners a practical Level 0 audit method. It is not a full AIx assessment, but it prepares learners to identify alignment risks in real systems.

11.1 What Is an Alignment Audit?

An alignment audit is a structured look at whether a system's behavior remains grounded in the constraints that determine its viability.

A Level 0 audit is plain language. It does not require advanced scoring. It asks learners to map the system, diagnose the proxy, inspect feedback, and recommend correction.

11.2 Audit Step 1: Define the System

A bad audit begins with a vague system. Be specific.

Instead of “healthcare,” audit “the appointment scheduling process in a local clinic.” Instead of “social media,” audit “the recommendation feed for short videos.” Instead of “work,” audit “the weekly reporting process for one department.”

Define:

- boundary: what is inside and outside the audit;
- purpose: what the system claims to do;
- actors: who participates;
- outputs: what the system produces;
- feedback: how the system learns.

11.3 Audit Step 2: Name the Stated Goal and Actual Objective

The stated goal is what the system says it serves. The actual objective is what the system rewards under pressure.

Examples:

- Stated goal: educate students. Actual objective: raise test scores.
- Stated goal: support customers. Actual objective: reduce call time.
- Stated goal: improve health. Actual objective: increase app streaks.
- Stated goal: provide accurate answers. Actual objective: satisfy user preference quickly.

11.4 Audit Step 3: Map the Layers

List the physical, biological, and constructed constraints.

Layer	Audit questions
Physical	What material, time, energy, factual, or ecological constraints apply?
Biological	What human needs, stressors, attention limits, trust conditions, or dignity issues apply?
Constructed	What rules, incentives, metrics, policies, roles, or agreements shape behavior?

11.5 Audit Step 4: Identify the Proxy

Identify the metric or visible signal that is treated as success.

Ask:

- What is easiest to measure?
- What appears on the dashboard?
- What gets rewarded?
- What do people protect under pressure?
- What number could improve while reality gets worse?

11.6 Audit Step 5: Find Hidden or Ignored Constraints

Hidden constraints often show up through recurring friction.

Look for:

- repeated complaints;
- unexplained turnover;
- quality problems;
- workarounds;
- quiet resentment;
- downstream rework;
- user regret;
- evidence gaps;
- safety concerns;
- trust erosion.

11.7 Audit Step 6: Evaluate Feedback

Ask whether the system can see, believe, and act on real consequences.

1. Who can report harm?
2. Are reports punished or welcomed?
3. Does feedback reach decision-makers?
4. How long does feedback take?
5. What feedback is missing from the dashboard?
6. Can the system correct after feedback arrives?

11.8 Audit Step 7: Diagnose Alignment Risk

Use this Level 0 risk scale:

Risk level	Description
Low	Metrics are grounded, feedback is real, and correction is easy.
Moderate	Some proxy risk exists, but the system can detect and correct problems.
High	Visible success may hide lower-layer cost; feedback is weak or delayed.
Severe	The system is strongly optimizing a proxy while suppressing feedback and delaying correction.

11.9 Audit Step 8: Recommend Correction

A good recommendation is specific and tied to the diagnosis.

Types of recommendations:

- add a counter-metric;
- create a feedback channel for affected people;
- slow down a high-risk process;
- revise incentives;
- add verification before output;
- protect dissent;
- make harms visible;
- create an appeal or rollback path;
- run a pilot before scaling;
- define a stop condition.

11.10 The Level 0 Audit Template

Basic Alignment Audit Template

1. System name and boundary.
2. Stated goal.
3. Actual rewarded objective.
4. Physical constraints.
5. Biological constraints.
6. Constructed constraints.
7. Main proxy.
8. Proxy failure mode.
9. Hidden or ignored constraints.
10. Feedback strengths.
11. Feedback weaknesses.
12. Alignment risk level.
13. Recommended correction.
14. First experiment or next step.

11.11 Worked Example: Remote Work Policy

System: a company remote work policy.

Stated goal: maintain collaboration and productivity.

Actual rewarded objective: visible availability during office hours.

Physical constraints: time zones, commuting time, home workspace, energy.

Biological constraints: focus, trust, belonging, autonomy, fatigue.

Constructed constraints: meeting norms, manager expectations, attendance tools, promotion criteria.

Main proxy: online presence and meeting attendance.

Proxy failure mode: people appear available while deep work and trust decline.

Hidden constraint: uninterrupted focus time.

Feedback weakness: workers fear appearing uncommitted if they report meeting overload.

Risk: high if presence is treated as productivity.

Correction: define output-based expectations, meeting budget, protected focus blocks, periodic trust survey, escalation path for overload.

11.12 Practice: Prepare for Capstone

Choose your capstone system and complete the audit template. Keep the system small enough to analyze but important enough to matter.

Good capstone choices:

- a workplace metric;
- a school policy;
- a personal habit system;
- an AI assistant use case;
- a social media feed;
- a hospital or clinic process;
- a public service workflow;
- a team communication routine.

Knowledge Check

1. Why must an audit define the system boundary?
2. What is the difference between stated goal and actual objective?
3. What should you look for when identifying hidden constraints?
4. Name the four Level 0 risk levels.
5. What makes a recommendation useful?

11.13 Certification Artifact

Submit a complete Level 0 Basic Alignment Audit using the template above.

MODULE 12

Certification Assessment

Module Aim

This module explains how learners earn the Certified Alignment Foundations credential and how instructors or reviewers evaluate the final work.

12.1 Credential Philosophy

The Level 0 credential is not awarded for memorizing slogans. It is awarded for demonstrating alignment literacy in a real or realistic system.

A certified learner should be able to notice when a system is confusing proxies with reality, identify the layers of constraint, inspect feedback, and propose a correction loop.

12.2 Assessment Components

The certification has four parts:

1. Chapter knowledge checks.
2. Learning artifacts from selected modules.
3. Final Basic Alignment Audit.
4. Short reflection on correction and uncertainty.

12.3 Passing Standard

A learner passes when they show competency in five areas:

Competency	Passing evidence
Layer recognition	Correctly distinguishes physical, biological, and constructed constraints.

Proxy diagnosis	Identifies a proxy and explains how it may drift from reality.
Feedback literacy	Describes what feedback is missing, delayed, distorted, or punished.
Correction thinking	Recommends at least one realistic correction loop.
Systems humility	Acknowledges uncertainty, tradeoffs, and limits of the audit.

12.4 The Final Capstone

The final capstone is a short Basic Alignment Audit of a chosen system. The preferred length is 1,500 to 2,500 words for written submission or 8 to 12 slides for presentation.

The capstone must include:

1. System boundary.
2. Stated goal and actual rewarded objective.
3. Layered constraint map.
4. Proxy diagnosis.
5. Feedback diagnosis.
6. Invisible divergence risk.
7. Alignment risk level.
8. Correction plan.
9. Reflection on uncertainty.

12.5 Capstone Examples

Acceptable capstone topics include:

- a workplace productivity dashboard;
- a classroom grading policy;
- a social media platform feature;
- a personal habit tracking app;
- an AI writing assistant workflow;
- a hospital discharge process;
- a hiring pipeline;
- a city's parking policy;
- a nonprofit donor reporting process;
- a customer support metric.

12.6 The Oral Defense

For live certification, learners give a 5-minute defense. Reviewers ask:

- What is the strongest evidence for your diagnosis?
- What might you be missing?
- Who would disagree with your audit and why?
- What feedback would change your mind?
- What correction would you try first?

The defense is not a debate. It is a test of humility, grounding, and correction readiness.

12.7 Scoring Overview

Each capstone is scored from 0 to 20:

Category	Points	What reviewers look for
System definition	2	Clear boundary and purpose
Layer map	4	Accurate P/B/C distinction
Proxy diagnosis	3	Clear target-proxy gap
Feedback diagnosis	3	Missing and distorted feedback named
Correction plan	4	Practical, specific, tied to diagnosis
Humility and tradeoffs	2	Acknowledges uncertainty and limits
Clarity	2	Coherent writing or presentation

Passing threshold: 14 out of 20.

Distinction threshold: 18 out of 20.

12.8 Ethical Use of the Credential

The Certified Alignment Foundations credential does not make someone an expert auditor of high-stakes systems. It certifies foundational literacy.

A certified learner should not use this credential to claim final authority over healthcare, legal, engineering, safety, or policy decisions. Instead, the credential shows that the learner can ask better questions, structure better conversations, and recognize common alignment risks.

12.9 Final Reflection

Learners complete a short reflection:

1. What proxy do you personally trust less after this course?
2. What feedback do you need to listen to more honestly?
3. What correction loop will you build in your work or life?
4. What uncertainty remains in your capstone audit?

Graduation Statement

A certified alignment-literate learner does not claim to see everything. They claim to know that no system sees everything, and that reality contact, feedback, and correction must therefore be designed and practiced.

MODULE A

Course Schedules

A.1 8-Hour Intensive Format

Time	Session
0:00–0:45	Orientation and the alignment lens
0:45–1:30	Why systems fail
1:30–2:30	Layered reality workshop
2:30–3:15	Reality vs proxies
3:15–4:00	Feedback and correction
4:00–4:45	Optimization pressure
4:45–5:30	Invisible divergence
5:30–6:15	Participation loop practice
6:15–7:15	Basic alignment audit lab
7:15–8:00	Capstone presentations and certification review

A.2 10-Hour Standard Format

Two 5-hour sessions.

Day 1

Alignment lens, systems failure, layered reality, proxies, first artifact.

Day 2

Feedback and correction, optimization, invisible divergence, participation loop, audit lab, assessment.

A.3 12-Hour Academy Format

Six 2-hour sessions.

Session	Focus
1	The alignment lens and why systems fail
2	Layered reality and constraint classification
3	Proxies, metric theater, and dashboard failure
4	Feedback, correction, and optimization pressure
5	Invisible divergence and everyday alignment
6	Capstone audit workshop and certification defense

MODULE B

Facilitator Guide

B.1 Teaching Posture

Teach this course as a practice of perception, not as ideology. The goal is to help learners see systems more clearly. Avoid turning alignment language into a new scoreboard for superiority.

B.2 Recommended Facilitation Norms

- Use examples before definitions.
- Invite learners to challenge the framework with cases.
- Keep the course non-mathematical.
- Distinguish diagnosis from blame.
- Keep high-stakes personal examples voluntary.
- Require humility in every capstone.

B.3 Discussion Prompts

1. What is a metric you once trusted more than you do now?
2. Where have you seen compliance replace purpose?
3. When did feedback arrive too late?
4. What constructed pressure feels physical in your life or workplace?
5. What does your organization punish people for saying?

B.4 Handling Resistance

Learners may resist the framework because it threatens existing success metrics. Do not force agreement. Ask them to test the framework:

- What would you expect to see if the system were misaligned?
- What feedback would prove the current metric is sufficient?
- Who would know before leadership knows?
- What would count as reality pushing back?

B.5 Group Exercise: Proxy Court

Divide learners into three roles:

- Proxy defender: argues why the metric is useful.
- Reality witness: describes what the metric misses.
- Correction designer: proposes a safer metric set.

Use grades, engagement, revenue, response time, or attendance as cases.

MODULE C

Quiz Bank

C.1 Multiple Choice

1. Which statement best defines alignment in this course?
 - A. Following all rules exactly.
 - B. Improving the main metric as fast as possible.
 - C. Maintaining coherence with real constraints over time.
 - D. Avoiding all mistakes.
2. A proxy becomes dangerous when:
 - A. it is measured.
 - B. it is treated as the reality it was meant to represent.
 - C. it is displayed on a dashboard.
 - D. it is qualitative.
3. Which is primarily a constructed constraint?
 - A. Gravity.
 - B. Sleep need.
 - C. A job title.
 - D. The finite number of hours in a day.
4. Invisible divergence occurs when:
 - A. a system fails visibly and immediately.
 - B. capability or performance improves while true alignment degrades.
 - C. no one measures anything.
 - D. a system refuses to optimize.
5. Correction capacity means:
 - A. the ability to avoid all errors.

- B. the ability to detect, revise, recover, or re-ground after error.
- C. the strength of the main metric.
- D. the number of people in the system.

C.2 Short Answer

1. Give one example of a physical, biological, and constructed constraint in the same system.
2. Explain the difference between feedback and data.
3. Describe a system where local rationality can produce global failure.
4. What is alignment debt, in plain language?
5. Why should proposals be treated as hypotheses?

C.3 Answer Key

Multiple choice: 1-C, 2-B, 3-C, 4-B, 5-B.

MODULE D

Capstone Rubric

D.1 Rubric

Criterion	Points	Distinguished performance
System boundary	2	System is specific, bounded, and understandable.
Stated vs actual objective	2	Clearly distinguishes mission from rewarded behavior.
Layer map	4	Accurately identifies physical, biological, and constructed constraints and at least one layer confusion.
Proxy diagnosis	3	Explains how the proxy can improve while reality degrades.
Feedback diagnosis	3	Identifies what feedback is missing, distorted, delayed, or punished.
Invisible divergence risk	2	Names a plausible pattern where internal success masks degradation.
Correction plan	4	Recommends specific, realistic correction actions tied to the diagnosis.
Uncertainty and tradeoffs	2	States what the audit might be missing and what evidence would change the conclusion.
Clarity	2	Communicates in plain language with coherent structure.

D.2 Reviewer Notes

Do not reward jargon. Reward clear sight. A learner who uses simple language to identify a real proxy failure should score higher than a learner who repeats terminology without diagnosis.

MODULE E

Certificate Template

The Alignment Science Academy

Certificate of Completion

Certified Alignment Foundations

This certifies that

[Learner Name]

has completed the Level 0 Alignment Literacy Certification and demonstrated foundational competence in layered constraints, proxy diagnosis, feedback integrity, correction thinking, and basic alignment audit practice.

Date:

Instructor:

MODULE F

Sources and Further Study

This Level 0 curriculum is derived from the broader ATS, AANA, and AIx research program and translated into a non-mathematical certification format.

Bibliography

- [1] Armando Sori. *The Fundamentals of Reality: On Participation, Meaning, and the Architecture of Existence*. Working manuscript, 2026.
- [2] Armando Sori. *A Theory of Alignment in Layered Systems (ATS 4.0)*. Working manuscript, 2026.
- [3] Armando Sori. *Alignment-Aware Neural Architectures: Constraint-Coherent Learning via Verifier-Grounded Correction*. Working manuscript, 2026.
- [4] Armando Sori. *Alignment Index (AIx): A Constraint-Based Measure of System Alignment Across Layered Reality*. Working manuscript, 2026.
- [5] Charles Goodhart. *Problems of Monetary Management: The U.K. Experience*. Papers in Monetary Economics, 1975.
- [6] Donella Meadows. *Thinking in Systems: A Primer*. Chelsea Green Publishing, 2008.
- [7] Peter Senge. *The Fifth Discipline: The Art and Practice of the Learning Organization*. Doubleday, 1990.